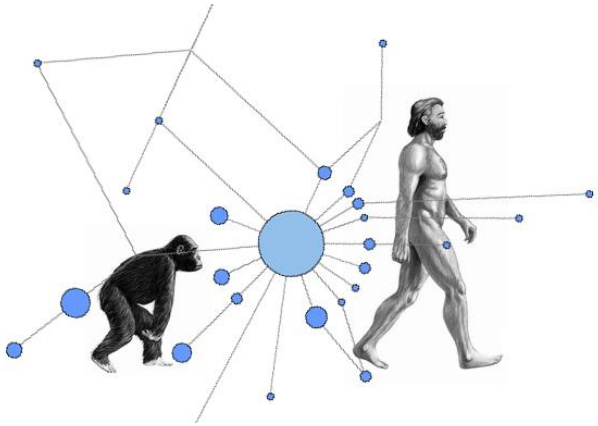


The human Whey-acidic-protein Four-Disulfide Core-domain (*WFDC*) cluster on 20q13 region: evolutionary history and role in human health and disease.

Zélia Alexandra Mota Quintas Ferreira
Biodiversity, Genetics and Evolution
Department of Biology
2013

Supervisor
Susana Seixas, PhD, IPATIMUP
Cosupervisor
Belen Hurle, PhD, NHGRI/NIH

Dissertação apresentada à Faculdade de Ciências da Universidade do Porto para obtenção do grau de Doutor em Biodiversidade, Genética e Evolução.

Nota prévia

Nos termos do nº 2 do artigo 8º do Decreto-Lei nº 388/70, foram incluídos em alguns capítulos desta dissertação os resultados de trabalhos publicados ou em publicação. Em todos estes trabalhos, a candidata participou na obtenção, análise e discussão dos resultados, bem como na elaboração da publicação.

Assim, os seguintes artigos foram publicados como parte da Dissertação:

Ferreira Z. *et al.* (2011) Differing evolutionary histories of *WFDC8* (short-term balancing) in Europeans and *SPINT4* (incomplete selective sweep) in Africans. Mol Biol Evol. 28(10):2811-22.

Ferreira Z. *et al*. (2013) Reproduction and Immunity Driven Natural Selection in the Human *WFDC* Locus. Mol Biol Evol. 30(4):938-50.

Ferreira Z. *et al*. Sequence diversity of *Pan troglodytes* subspecies and the impact of *WFDC6* selective constraints in reproductive immunity. (submetido)

Ferreira Z. *et al*. Characterization of the Human *WFDC8*: Evolutionary history and differential allele expression. (em preparação)

A instituição de origem da candidata foi a Faculdade de Ciências da Universidade do Porto (FCUP), tendo o trabalho sido realizado sob orientação da Doutora Susana Seixas (Instituto de Patologia e Imunologia Molecular da Universidade do Porto - IPATIMUP), Doutora Belen Hurle (*National Human Genome Research Institute / National Institutes of Health* – NHGRI/NIH), e Professor Doutor Andrew G. Clark (*Cornell University*).

Este trabalho foi parcialmente apoiado pelo Intramural Research Program do NHGRI/NIH; e pela Fundação para a Ciência e a Tecnologia comparticipado pelo Fundo Social e Fundo Comunitário Europeus, e por fundos nacionais do Ministério da Educação e Ciência (POPH - QREN - Promoção de Emprego Cientifico e COMPETE – Programa Operacional Temático Factores de Competitividade) através da atribuição de uma bolsa individual de doutoramento (SFRH / BD / 45907 / 2008) e através de dois projectos de investigação (PTDC/SAU-GMG/64043/2006 e PTDC/BEX-GMG/0242/2012).

“*We came in the wind of the carnival.*
A wind of change, or promises.
The merry wind, the magical wind, making March hares of everyone, tumbling
blossoms and coat-tails and hats;
rushing towards summer in a frenzy of exuberance.”

Joanne Harris

Acknowledgments/Agradecimentos

This dissertation would not have been possible without the collaboration, input, and encouragement of a *lot* of people. During these four years the experiences shared between all of the people involved will always be remembered with appreciation.

Gostaria de agradacer a minha orientadora, Doutora Susana Seixas, que apostou em mim em 2007 quando era ainda uma aluna sem experiência. Agradeço ter-me orientado durante todo este tempo, sempre paciente, e por acreditar em mim e me ter incentivado a concorrer a uma bolsa de doutoramento. Este percurso não teria sido possível sem ela, pelo que agradeço os momentos de discussão científica, cafés, conferências e gchat que passamos. Mesmo longe, foi um pilar no meu desenvolvimento cientifico e ficarei para sempre grata por ter sido sua aluna.

To my co-supervisor, Dr. Belen Hurle, who gave me the opportunity to join one of the greatest laboratories in the world at the NHGRI, and to include her exciting project in my dissertation. She has guided me scientifically, taught me how to weave myself in the scientific world, and helped me discover various ways to succeed as a scientist, and personally. I will be always grateful for the great advice she gave me and I will always be gratefull for having been her student.

To my mentor Professor Andy Clark, for giving me the great opportunity to visit his laboratory and to interact with him and his group. I deeply appreciate the scientific discussions, the lab meetings, and for sharing his enthusiasm on human evolution. I deeply appreciate the feedback he gave on this thesis, which helped me improve my knowledge of population genetics.

Á Fundação para a Ciência e a Tecnologia, agradeço a concessão da bolsa de doutoramento, SFRH / BD / 45907 / 2008 sem a qual a realização deste trabalho não teria sido possível.

Aos Professores Doutores Jorge Rocha, Paulo Alexandrino e Nuno Ferrand, ao Programa Doutoral em Biodiversidade, Genética e Evolução, e á Faculdade de Ciências da Universidade do Porto por me terem dado a oportunidade de seguir para Doutoramento, por todos os conselhos e por terem criado um bom ambiente cientifico de forma a que podesse iniciar o meu percurso.

Ao Professor Doutor Sobrinho Simões agradeço a oportunidade de integrar uma equipa de investigadores notável como a que encontrei no IPATIMUP.

I would like to thank Dr. Eric Green, Director of NHGRI/NIH, for giving me the opportunity to integrate into the greatest research institution, for reading my manuscripts when his life was so busy already, and for the constructive criticism that allowed me to grow as a scientist during my time there.

I would also like to thank Dr. Larry Brody, director of the GTB (NHGRI/NIH) for taking Belen and I under his wing, mentoring me through the design of new experiments and helping me build my scientific network.

Gostaria também de agradecer ao Banco de Tumores e Tecidos do Hospital de São João, Professora Doutora Fátima Carneiro e Dr. Carla Bartosch, pelo fornecimento de amostras sem as quais algumas experiências não teriam sido possíveis.

I would like to acknowledge all of the co-authors of the publications presented in this thesis for their contributions. A special thanks to Aida Andrés for all her scientific input on all the manuscripts written.

Aos correntes e antigos membros do grupo *Genetics, Evolution and Pathology /Evolutionary Funcional Genomics*/*Proteolysis in Diseases*, um enorme obrigada pela forma que me apoiaram desde que integrei no grupo. Pela partilha de experiencias (dentro e fora do laboratório) e momentos de bom humor que foram passados durante o meu doutoramento. Obrigada a vocês, que tornaram os momentos no laboratório 108 inesquecíveis: Margarida, Isabel, Sandra, Sílvia e Paula.

Um especial obrigada a Patrícia, que me apoiou sempre e que tornou o "aquário" uma casa para mim quando regressei. Patrícia, a amizade desenvolvida ficará para sempre guardada, e quando acabares, estarei a teu lado para as "pausas" para manteres a sanidade mental como fizeste comigo (nem que seja por Skype!). E não me esqueci da nossa Review!

To all the members of the Green lab, that made my presence in the laboratory not only a very exciting scientific experience, but also a great two years that all

students deserve to have. A special thanks to Arjun and Winni for the scientific input, code hours and friendship.

To all the members of the Clark lab, who made six months in Ithaca go by so fast, for their scientific input, for their support and for making me feel at home in such a short period of time. A special thanks to Margarida, Sri, Haley, and Julia for being great lab mates and friends.

Agradeço todos os colegas do IPATIMUP pelo bom ambiente nos corredores e pelas técnicas que me ensinaram durante o meu doutoramento, em particular à Vânia e à Rita pela ajuda nos ensaios de luciferase, e à Andreia e à Cristina. Um especial agradecimento à Cristiana, que partilha comigo as aventuras nos Estados Unidos e cujo apoio, humor, e amizade fizeram este periodo divertido e mais "light".

Aos meus amigos no Porto, que tiveram de se despedir de mim (várias vezes), que sempre acreditaram e me apoiaram todas as minhas decisões mesmo quando isso significou enviarem-me para o outro lado do Oceano. Joana, Ana Lua, Gi, Susana, Alonso, Ké, e João, sem vocês do meu lado este doutoramento não teria sido possível. Sem a vossa amizade nada disto faz sentido.

To my friends Nadine, Dan, Michelle, Nick, Kristin, Jessie, Petra and all the others I don't have enough space to mention… you guys made 'America my home. I'll see y'all soon! An enormous *THANK YOU* to Karlye, for making me believe I can do this, for giving me strength when I didn't have it, and for being the most amazing woman I know. I could not have done it without you.

To Steve, for being my pillar and my home no matter where I am. For helping me write letters, for dealing with my dissertation stress, and for making my life *that* much better. The end of this phase of my life is down the road from us, I look forward to seeing what comes next with you by my side.

Aos meus familiares, pelo apoio que sempre me deram e pelas palavras de encorajamento. Á minha irmã, Mafalda, que com a sua alegria me fez sentir em casa dentro e fora de Portugal. Aos meus pais, por todo o seu apoio incondicional. E aos meus avós, que apesar de não entenderem bem o que é esta coisa da Biologia, estão sempre orgulhosos de mim.

Summary

The development of large-scale catalogs of genetic variation has stimulated interest in finding targets of positive selection, leading to the development of a plethora of genome-wide scans for positive selection. The identification of these genes can prime insights into how individuals are predisposed to disease, and may inform the development of improved therapeutic and disease-prevention strategies. Some gene ontology categories are enriched for signals of positive selection, among which are genes involved in immunity and host-pathogen interactions and reproduction. In this regard, a genome-wide scan for positive selection has pinpointed several Whey-Acidic-Protein (WAP) four-disulfide core domain (*WFDC*) genes to be under positive selection in human populations.

WFDC genes encode 17 serine protease inhibitors, which in most cases are not functionally characterized, but well-studied *WFDC* genes have biological functions associated with innate immune response and reproduction in mammals. Additionally, neighboring genes encode seminal proteins Semenogelin 1 and 2 (SEMG1 and SEMG2), the main components of the semen coagulum. *WFDC* and *SEMG* genes have demonstrated a strikingly high rate of amino acid replacement (d_N/d_S), indicative of responses to adaptive pressures during vertebrate evolution. In addition, *SEMG2* evolution was previously shown to be correlated with promiscuity rates in primates.

The complex and rapid evolution *of WFDC* and *SEMG* in primates motivated the characterization of their genetic variation and the extent to which natural selection has shaped their evolution in human populations. The evolution of *WFDC* and *SEMG* genes in chimpanzees was also pursued in an effort to contrast selective pressures in two closely related species. The main goal was to identify which *WFDC* and *SEMG* genes are evolving under stronger selective pressures and should be further assessed for their biological function.

The study was designed in two stages: the first was based on a genome-wide scan for positive selection and the second enclosed a high-throughput comprehensive sequencing study of the *WFDC* locus. A signature of short-term balancing selection in *WFDC8* in Europeans was identified, characterized by an intermediate-frequency long haplotype linked to the proposed candidate variant -44A. In addition, a signal of recent positive selection (incomplete selective sweep) in Africans was pinpointed as the result of a high frequency haplotype defined by the Ser73+98A configuration. In the second stage, a follow-up approach was taken to perform a systematic high-throughput Sanger sequencing effort across the entire *WFDC* locus in the three major HapMap human populations, and in three chimpanzee subspecies (*P. t. troglodytes*, *P. t. ellioti* and *P. t.*

verus). The comprehensive analysis of the human genetic variation allowed confirmation of the previously identified signals, and endorsed the identification of a signature of selection on standing variation centered in *SEMG1* in Asians. The candidate variant was hypothesized to be Thr56Ser, which may alter the proteolytic profile of SEMG1 and antimicrobial activities of semen. The heterogeneous signals of non-neutral evolution of *WFDC* and *SEMG*s in human populations led to propose that they are primary targets of selection, likely due to immune response to pathogens and reproductive pressures.

Conversely, when genetic variation in chimpanzee subspecies was analyzed the signals of selection pointed toward purifying selection acting in *WFDC6* and *EPPIN*, genes involved in antimicrobial defense in the reproductive tract. Multispecies studies coupled with analyses of summary statistics suggested *WFDC6* is evolving to be functionally divergent from *EPPIN* and to target different proteases and to increase the response against pathogens. Due to the function of these genes and the promiscuous mating patterns of chimpanzees, it is hypothesized that the selective pressures shaping the genetic diversity in chimpanzees are driven by increased protection from sexually transmitted diseases.

Overall, it was confirmed that the *WFDC* genes are evolving under different pressures in humans and chimpanzees, with selection targeting different genes in the two species. The time scale of the selective pressures was found to be more recent in humans.

Three genes were prioritized for future functional characterization: *WFDC8*, *SPINT4* and *SEMG1*. For *WFDC8*, preliminary *in vitro* and *in vivo* studies were inconclusive in clarifying the functional impact of the candidate variant -44A in gene expression/regulation. Specifically, the effect of this variant in the expression of *WFDC8* appears to be dependent on the tissue background, where -44A enhances expression in ovaries (*in vivo*) and HeLa cell lines (*in vitro*); and decreases expression in Caco-2 cell lines (*in vitro*). Further assessment of WFDC8 functionality as an antimicrobial and protease inhibitor protein can elucidate not only the consequences of the short-term balancing selection event but also the role WFDC8 plays in innate immunity in the reproductive tract. The deep characterization of other genes in the *WFDC* locus will increase the knowledge on the biological dynamics of rapidly evolving genomic regions in primates.

Sumário

O desenvolvimento de novos catálogos de variação genética tem vindo a estimular o interesse em encontrar alvos de seleção positiva, levando ao desenvolvimento de uma panóplia de *scans* genómicos para deteção de seleção positiva. A identificação de genes envolvidos na seleção positiva poderá ajudar a compreender a predisposição de indivíduos a certas doenças e bem como orientar no desenvolvimento de estratégias de terapia e prevenção de doenças. Algumas categorias génicas são mais enriquecidas em sinais de seleção entre as quais genes envolvidos em imunidade, interações hospedeiro-patogénio e reprodução. Neste contexto, um *scan* genómico para a deteção de seleção positiva identificou uma vantagem seletiva em vários genes de *Whey-Acidic-Protein* (*WAP*) *four-disulfide core domain* (*WFDC*) em populações humanas.

Os genes *WFDC* codificam 17 inibidores de protease de serina que na sua maioria, não estão ainda caracterizados funcionalmente, apesar de os genes *WFDC* com função biológica conhecida estarem associados com a resposta imune inata e a reprodução em mamíferos. Os genes vizinhos aos *WFDC* codificam as proteínas Semenogelinas 1 e 2 (SEMG1 e SEMG2), que são as principais proteínas do coágulo seminal. Os genes *WFDC* e *SEMGs* demonstraram um elevado rácio de substituições amino acídicas (d_N/d_S) indicando uma resposta a pressões adaptativas durante a evolução de vertebrados. Por exemplo, a evolução acelerada da *SEMG2* foi previamente correlacionada com os níveis de promiscuidade em primatas.

A complexa e rápida evolução dos *WFDC* e *SEMG* em primatas encorajou a caracterização da variação genética destes genes, assim como o estudo da seleção natural que tem vindo a influenciar a sua evolução em populações humanas. A evolução dos genes *WFDC* e *SEMG* em chimpanzés foi também analisada de forma a contrastar as pressões seletivas em duas espécies próximas. O principal objetivo foi identificar quais genes *WFDC* e *SEMG* estão a evoluir sobre fortes pressões seletivas e deverão ser caracterizados biologicamente.

Num estudo desenhado em duas fases, a primeira foi baseada num estudo genómico de deteção de seleção positiva e a segunda baseada num estudo sistemático de todos os genes *WFDC*. A primeira fase permitu identificar uma assinatura de seleção balanceadora recente no gene *WFDC8* em Europeus, caracterizada por um longo haplótipo de frequência intermédia associado ao variante candidato -44A. Este mesmo estudo identificou ainda um sinal em Africanos de seleção positiva recente (*incomplete selective sweep*) associado a uma configuração haplotipica de elevada frequência definida pelas variantes Ser73+98A. A segunda

parte do estudo envolveu a sequenciação Sanger de todo a região *WFDC* em três populações humanas do HapMap e em três subespécies de chimpanzés (*P. t. troglodytes*, *P. t. ellioti* e *P. t. verus*). Este estudo extensivo da variação genética humana permitiu a confirmação dos dois sinais identificados previamente, e proporcionou a identificação de um sinal de seleção sobre variação pré-existente, centrada no gene *SEMG1* em Asiáticos. O variante candidato proposto, Thr56Ser, poderá alterar o perfil proteolítico da SEMG1 e as atividades antimicrobianas do sémen. Baseado na heterogeneidade de sinais de evolução não-neutra dos genes de *WFDC* e *SEMGs* em populações humanas, os resultados deste trabalho sugerem que estes genes são alvos primários de seleção positiva, provavelmente devido a pressões ligadas à resposta imune a patogénios e a reprodução.

Por outro lado, quando a variação genética das subespécies de chimpanzés foi analisada, os sinais de seleção apontaram para um sinal de seleção purificante centrado nos genes *WFDC6* e *EPPIN,* cujas proteínas estão envolvidas na defesa antimicrobiana do aparelho reprodutor. Estudos comparativos de várias espécies combinados com estatísticas sumárias sugeriram que o gene *WFDC6* está a evoluir de forma a se tornar funcionalmente divergente da EPPIN para interagir com diferentes proteases e para aumentar a resposta contra os patogénios. Devido à função destes genes e à natureza promiscua dos chimpanzés, propôs-se que as pressões seletivas que estão a moldar a diversidade genética em chimpanzés são desencadeadas pelo aumento da proteção contra doenças sexualmente transmissíveis.

Em resumo, confirmou-se que os genes *WFDC* estão a evoluir sobre pressões seletivas divergentes em humanos e chimpanzés, com eventos de seleção natural a ocorrer em genes diferentes nas duas espécies. Os resultados apresentados neste trabalho indicam ainda que as pressões seletivas em humanos parecem ser bastante mais recentes que em chimpanzés.

Três genes foram determinados como prioritários para uma caraterização funcional futura: *WFDC8*, *SPINT4* e *SEMG1*. No caso do gene *WFDC8*, os resultados preliminares *in vitro* e *in vivo* foram inconclusivos na clarificação do impacto funcional do variante candidato -44A na regulação/expressão do gene. O efeito desta variante na expressão de *WFDC8* parece depender do tecido estudado, uma vez que o variante -44A aumenta a expressão do *WFDC8* em ovários (*in vivo*) e na linha celular de HeLa (*in vitro*), mas diminui a expressão na linha celular Caco-2 (*in vitro*). Mais estudos do *WFDC8* deverão ser efetuados para determinar a sua atividade antimicrobiana e de inibição proteolítica de forma a elucidar as repercussões da seleção balanceadora recente e o papel do WFDC8 na imunidade inata do aparelho reprodutivo. A caracterização exaustiva dos outros genes sobre seleção positiva no

locus *WFDC* irá potenciar o conhecimento na dinâmica biológica de regiões que evoluem de forma acelerada em primatas.

Abbreviation list

ADAM1 - a disintegrin and metalloproteinase domain 1

ADAM2 - a disintegrin and metalloproteinase domain 2

AGT - angiotensinogen (serpin peptidase inhibitor, clade A, member 8)

Af - Africans

BEB - Bayes empirical Bayes

bp - basepair

Caco-2 - Colorectal cancer cell line

cDNA – complementary DNA

CEU – Utah residents with Northern and Western European ancestry from the CEPH collection

CHB - Han Chinese in Beijing, China

CLU – clusterin

CYP3A - Cytochrome P450, Family 3, Subfamily A

DARC - Duffy blood group, chemokine receptor

DIND – Derived intra-allelic nucleotide diversity

DMEM - Dulbecco's Modified Eagle Medium

DNA – Deoxyribonucleic acid

EHH - Extended Haplotype Homozygozity

EPPIN – Epididymal protease inhibitor (also known as, *SPINLW1* and *WFDC7*)

eQTLs - expression Quantitative Trait Loci

EU - Europeans

f – allele frequency

FIN - HapMap Finnish individuals from Finland

G6PD - glucose-6-phosphate dehydrogenase

GAGs - glycosaminoglycans

GAPDH - Glyceraldehyde 3-phosphate dehydrogenase

gDNA - Genomic DNA

GBR - British individuals from England and Scotland

GO – gene ontology

GWAS - Genome-Wide Association Studies

GWS - Genome-Wide Scans

HBB – hemoglobin beta

HbS – hemoglobin S

HE4 - Human epididymis protein 4 (also known as WFDC2)

HeLa - Human Cervix Adenocarcinoma cell line

HIV – Human Immunodeficiency Virus

HKA – Hudson-Kreitman-Aguadé

HWE - Hardy-Weinberg equilibrium

IBS - Iberian populations in Spain

iHS – integrated Haplotype Score

Indels - insertions/deletions

JPT - Japanese in Tokyo, Japan

KLK3 – Kallikrein-3 (also known as *PSA*)

KYA – thousand years ago

LCT - lactase

LD - Linkage Disequilibrium

LPS – lipopolysaccharide

LRH – Long Range Haplotype

LTF - lactoferrin

Mya – million years ago

Mb – Megabases

MHC - Major Histocompatibility Complex

MKT - McDonald-Kreitman test

mRNA – messenger ribonucleic acid

MSMB - microseminoprotein beta

mtDNA - mitochondrial DNA

MW – Molecular Weight

MWUhigh – Mann-Whitney U high

MZF - myeloid zinc finger

NF-κB - nuclear factor κB

NSyn – Nonsynonymous

PC - Principal Component

PCA – Principal Component Analysis

PCR – Polymerase Chain Reaction

PI3 - peptidase inhibitor 3, also known as *elafin*

PPAR - peroxisome proliferator-activated receptor

PSA - prostate-specific antigen (also known as *KLK3*)

REHH – Relative Extended Haplotype Homozygozity

REST - Rapidly Evolving Substrates for Transaminases

RNA – ribonucleic acid

RT-PCR – reverse transcriptase polymerase chain reaction

SEMG1 – Semenogelin 1

SEMG2 - Semenogelin 2

SFS – Site Frequency Spectrum

SLC25A5 - solute carrier family 25 (mitochondrial carrier; adenine nucleotide translocator), member 5

SLPI - Secretory Leukocyte Peptidase Inhibitor

SNP – Single Nucleotide Polymorphism

SPAG6 - sperm associated antigen 6

SPINT3 - serine protease inhibitor, Kunitz type 3

SPINT4 - serine protease inhibitor, Kunitz type 4

SPINT5 - serine protease inhibitor, Kunitz type 5

STDs - Sexually Transmitted Diseases

SVS - seminal vesicle secreted proteins

SWAM1 - Wfdc15b WAP four-disulfide core domain 15B (Mus muscullus)

SWAM2 - Wfdc12 – WAP four-disulfide core domain 12 (*Mus muscullus*)

Syn - Synonymous

TGM4 - transglutaminase 4

TLRs - Toll-Like Receptors

TMRCA – Time of most recent common ancestor

TSI - Toscani in Italy

UCSC – University of California Santa Cruz

UTR – Untranslated region

WAP – Whey-Acidic-Protein

WFDC - Whey-acidic-protein Four Disulfide Core

WFDC-CEN – *WFDC* centromeric sublocus

WFDC-TEL – *WFDC* telomeric sublocus

YRI – Yoruban in Ibadan, Nigeria.

ZP3 - zona pellucida glycoprotein 3

Contents

Figures list

General Introduction

Article 1

Article 2

Article 3

Article 4

Final Discussion

Appendices

Appendix 1

Appendix 2

Appendix 3

Appendix 4

Tables list

Article 1

Article 2

Article 3

Article 4

Appendixes

Appendix 1

Appendix 2

Appendix 3

Appendix 4

1. General Introduction

1.1 Natural Selection in Humans

Natural selection, described by Darwin in 1872, is the principle in which beneficial traits make their carriers survive longer and reproduce more, tending to become more frequent in populations over time (Darwin 1859; Jobling, Hurles and Tyler-Smith 2004; Hartl and Clark 2007). Selection can occur at any stage of the development of an individual, from the zygote to the reproductive age at the moment of generating progeny, going through fertility and gamete selection, and fecundity and mate choice. Altogether, these factors contribute to the fitness of an individual in a particular environment where the important factor is the relative fitness of a phenotype compared to the other phenotypes competing for the same resources (Jobling, Hurles and Tyler-Smith 2004; Orr 2009).

Even though relative fitness is expressed as a function of phenotypic variation (such as height, weight, resistance to diseases, etc.), natural selection can only occur in a population if mutation has generated heritable genetic variation (polymorphisms). These mutations can be advantageous (adaptive), deleterious or neutral (having no effects on the fitness of an individual). A variant that increases the fitness of individuals will be spread to fixation (frequency = 1) by positive selection, whereas a deleterious variant will be eliminated by negative (or purifying) selection — a force to which all genes are likely subject to (Sabeti, et al. 2006; Kelley and Swanson 2008; Verrelli, et al. 2008; Akey 2009; Nielsen, et al. 2009; Barrett and Hoekstra 2011; Crisci, et al. 2011; Fumagalli, et al. 2011).

Some noncoding regions have been found to present hallmarks of purifying selection, highlighting not only the conservation of protein-coding regions (Bustamante, et al. 2005; Clark 2006; Clark, Aagaard and Swanson 2006; Blekhman, et al. 2008; Fu, et al. 2013; Moleirinho, et al. 2013), but also noncoding regions with important functional elements, such as microRNAs and regulatory elements (Bersaglieri, et al. 2004; Woolfe, et al. 2004; Siepel, et al. 2005; Quach, et al. 2009; Ward and Kellis 2012). Furthermore, other regions seem to exhibit patterns consistent with the action of positive and balancing selection (Nielsen, et al. 2005; Williamson, et al. 2005; Nielsen, et al. 2007; Akey 2009; Clark, et al. 2009; Raj, et al. 2012; Vernot, et al. 2012; Grossman, et al. 2013). Balancing selection favors diversity and therefore the selected variant never reaches fixation, maintaining two or more alleles at a single locus in a

population, and is more frequent than previously thought (Charlesworth 2006; Cagliani, et al. 2008; Andrés, et al. 2009; Fumagalli, et al. 2009; Andrés, et al. 2010; Campbell, et al. 2012; Bronson, et al. 2013; Leffler, et al. 2013).

Natural selection continually shapes the human genome and, with the availability of genome-scale genotyping data it has been possible to identify various regions that appear to have been targeted by selective pressures. The identification of these variants could lead to insights into how genes predispose individuals to disease, and could inform the development of improved therapeutic and disease-prevention strategies, as many of the selected genes are disease associated (Jobling, Hurles and Tyler-Smith 2004; International HapMap Consortium 2005; Nielsen 2005; Zhang, Nielsen and Yang 2005; Suk, et al. 2011). Selective events associated in indirect ways to diseases could result in antagonistic pleiotropy, a phenomenon in which adaptation also brings deleterious effects (Clark and Swanson 2005). Furthermore, as an advantageous allele spreads through a population during a selective sweep, alleles in LD with the advantageous allele will rapidly increase in frequency (genetic “hitchhiking”), which can also drive deleterious disease-causing alleles to high frequency (Seger, et al. 2010; Huff, et al. 2012).

It has been challenging to clearly identify evidence of selection in the human genome as variability in the genome is also strongly affected by demographic processes, such as population growth or bottlenecks; and by variation in the mutation and recombination processes. Nevertheless, there has been progress in identifying the role of population history, not only to help tease apart selective signals from demographic history but to answer the human curiosity as to our origins (Sabeti, et al. 2002; Schaffner, et al. 2005; Voight, et al. 2005; Gutenkunst, et al. 2009; Laval, et al. 2010; Gravel, et al. 2011). All the models agree that during the “Out of Africa” event, African and non-African populations diverged (over 50–100 KYA), and the environment and ecology of humans changed greatly. As humans spread out of Africa to the other continents, they experienced a vast range of new climates, diets and ecosystems. Humans also encountered new pathogens as they moved around the globe and moved into close proximity with domesticated animals and as human population densities increased (Akey 2009; Bustamante and Ramachandran 2009; Novembre and Di Rienzo 2009; Barreiro and Quintana-Murci 2010; Kim, et al. 2010; Pritchard and Di

Rienzo 2010; Gravel, et al. 2011; Hernandez, et al. 2011; O'Bleness, et al. 2012; Scally, et al. 2012; Tennessen, et al. 2012).

At the rate at which the world population is currently growing, there has been an explosive growth expanding the global population by at least three orders of magnitude over the past 400 generations, reaching 7 billion people. Even though efforts have been made to include this explosive growth, we should be careful when looking for signals of positive selection or when studying complex diseases as this implies massive departures from population genetics equilibrium generating potential false positives (Keinan and Clark 2012).

With the new sequencing techniques (The 1000 Genomes Project Consortium 2010; Neafsey and Haas 2011), studying complex diseases and the evolution of genomes has become easier, due to an ablility to sequence large numbers of individuals at lower cost. Some panels of genetic variability generated are enriched in common variants, such as the HapMap Project (International HapMap Consortium 2005; McVean, Spencer and Chaix 2005; Thorisson, et al. 2005; Frazer, et al. 2007; Altshuler, et al. 2010) while others are more focused on cataloging rare variants, such as “The 1000 Genomes Project” (f ~ 5%) (The 1000 Genomes Project Consortium 2010). Although none of these databases completely captures the human genetic variation (Voight, et al. 2006; Fu, et al. 2013), they are of great use in detecting footprints of selection, as shown by the plethora of genome-wide scans (GWS) that have uncovered selective signatures. Additionally, GWS for positive selection detect not only recent events based on the analysis in intraspecies polymorphism but also older events. Selective events over a long period of time can increase the fixation rate of advanatgous variants, and these changes can be detected by comparing DNA sequences between species (Bustamante, et al. 2005; Biswas and Akey 2006; Sabeti, et al. 2006; Sabeti, et al. 2007; Blekhman, et al. 2008; Kosiol, et al. 2008; Akey 2009; Kudaravalli, et al. 2009).

By comparing the rate of nonsynonymous changes with the rate of synonymous [d_N/d_S ratio; (Yang 2007)] or by comparison with intraspecies diversity [McDonald-Kreitman test; (McDonald and Kreitman 1991)] it is possible to detect old selective events (around 6 MYA). However, a more recent signal of selection (younger than 1 MYA) cannot be detected by multispecies comparisons. A more recent selective signal

is generally characterized by low levels of diversity, as a result of ‘‘hitchhiking’’, creating a ‘”selective sweep” (or hard sweep), which alters the typical pattern of genetic variation in the surrounding genomic region of an advantageous allele. Selective sweeps can be detected through the analysis of genetic variation within species, by using summary statistics such as Tajima’s *D*, Hudson-Kreitman-Aguadé (HKA), and Fu and Li’s *D** tests (Hudson, Kreitman and Aguadé 1987; Tajima 1989; Fu 1997). If, however, the selective sweep has occurred more recently, these tests will not significantly depart from neutrality. In a recent selective sweep derived alleles hitchhike to high frequency but may not reach fixation due to an incomplete (partial) sweep or recombination events, resulting in a region containing many high-frequency derived alleles. We can assess this excess by using Fay and Wu’s *H* test (Fay and Wu 1999; Zeng, et al. 2006). Recent signatures of positive selection also result in long-range haplotypes, creating homozygous haplotypes that are more extended the stronger or more recent the selective sweep is (Extended Haplotype Homozygozity – EHH) (Sabeti, et al. 2002; Voight, et al. 2006). Interestingly, these signals are most of the time population specific, as separate populations are subject to distinct environmental or cultural pressures. This adaptation to a particular environment will generate population specific frequencies, referred to as population differentiation that can be detected with the F_{ST} statistic (Excoffier 2002). Classical examples of recent selective sweeps include skin pigmentation, malaria resistance and lactose tolerance (Bersaglieri, et al. 2004; Coelho, et al. 2005; McEvoy, Beleza and Shriver 2006; Verrelli, et al. 2006; Voight, et al. 2006; Hancock, et al. 2011) .

Overall, studies have identified many genes showing strong evidence of positive selection and have demonstrated differences between populations and affected genomic regions. Certain gene ontology categories including sensory perception, dietary changes, immunity and host-pathogen interactions, reproduction and proteolysis were found to be enriched in gene under selective pressures (Bustamante, et al. 2005; Sabeti, et al. 2006; Sabeti, et al. 2007; Kosiol, et al. 2008; Akey 2009; Bustamante and Ramachandran 2009; Manry, et al. 2011; Grossman, et al. 2013). The most notable case comes from adaptation to dietary changes, where a polymorphism located upstream of the *LCT* gene was found to be under positive selection in Europeans, affecting *LCT* transcription levels and the ability of lactose to be digested throughout life (Bersaglieri, et al. 2004; Coelho, et al. 2005). Furthermore, African pastoralists (populations that rely on cattle domestication) are also able to

digest milk throughout life due to the convergent evolution of *LCT* gene (Tishkoff, et al. 2007). *LCT* is one the strongest signals of a partial selective sweep in genome-wide scans for selection in humans (Chimpanzee Sequencing and Analysis Consortium 2005; Voight, et al. 2006; Sabeti, et al. 2007; Williamson, et al. 2007).

Among the plethora of human selective events are genes with interconnected functionality in immune response and reproduction. For example, *TGM4* and *MSMB* play roles in suppressing immune response against sperm in the reproductive tract and they both show strong signs of positive selection (Clark and Swanson 2005). The innate immune response of an organism is the first barrier against invading microorganisms, providing immediate defense. It relies on pattern-recognition receptors, which recognize conserved and largely invariant microbial molecules, and trigger diverse mechanisms that initiate inflammatory and immune responses. They also activate an adaptive immune response, generating cascades such as phagocytosis, pro-inflammatory pathways, and the release of antimicrobial peptides. The best characterized innate immune genes are the *TLRs*, which are expressed on the cell surface and intracellular organelles and detect microorganisms and several bacterial products. Remarkably, intracellular *TLRs*, particularly specialized in viral recognition, have evolved under strong purifying selection whereas cell-surface *TLRs* have evolved under more relaxed selective constraints (Barreiro, et al. 2009).

In another example, one of the most studied cases of immunity related natural selection includes the disease resistance to malaria infection in Africa. Malaria is transmitted by the mosquito *Anopheles sp*., carrying either *Plasmodium falciparum*, the most deadly strain, or by *Plasmodium vivax*. Malaria prevalence and transmission increased as a consequence of the transition to agriculture, where the human populations changed from hunter-gatherers and a nomadic lifestyle to an agriculturally based settling. With the agricultural development, and cattle domestication, the levels of insects increased greatly, as insects gather preferably around animals and still water. Simultaneously, with the abundance of food populations increased exponentially in size, increasing contact between individuals. All of these factors contributed not only to the transmission of malaria but also other pathogens as they transferred from animals to humans (Sabeti, et al. 2002; Verrelli, et al. 2006; Mackinnon and Marsh 2010; Hedrick 2011; Rayner, et al. 2011). A number of common polymorphisms arose to confer resistance to malaria in African populations, some of them leading variants to

reach fixation (the Duffy null allele at the *DARC* gene), or to higher frequencies (such as HbS at the *HBB* gene and A- at the *G6PD* gene) (Akey, et al. 2004; Hancock, et al. 2008; Kelley and Swanson 2008).

Further cases of immunity related selection are the MHC class I and class II molecules, as one of the few thoroughly described cases of balancing selection The *MHC* locus is extremely polymorphic and some of its ancestral polymorphisms have been maintained for millions of years; in addition, humans and chimpanzees share 11 nonsynonymous *MHC* polymorphisms showing one of the rarest and oldest balancing selection signals (Takahata 1993; Hughes and Yeager 1998; Klein, et al. 1998; Prugnolle, et al. 2005; Fumagalli, et al. 2009; Garamszegi and Nunn 2011; Leffler, et al. 2013). As the MHC molecules determine the probability with which a given pathogen will be recognized by the individual's immune system, this extreme variability ensures the detection of many different pathogens, improving the effectiveness of the immune system, which is essential to the survival of hosts in the constantly changing pathogen spectra (Hughes and Yeager 1998; Klein, et al. 1998; Prugnolle, et al. 2005; Fumagalli, et al. 2009; Abi-Rached, et al. 2010; Garamszegi and Nunn 2011).

Signals of positive selection and rapid evolution have also been identified in various taxa related to reproduction and fertility (Swanson and Vacquier 2002). These genes are involved in gamete recognition, seminal fluid factors, and proteins in the male and female reproductive tract. One example of a signature of selection has been found in European Americans and Han Chinese for *SPAG6*, a gene involved in sperm motility (Williamson, et al. 2007). Another type of selection related to reproduction that has been found in various primate species is post-copulatory selection. Post-copulatory sexual selection encompasses both male–male competition in the form of sperm competition, and cryptic (hidden) female choice. Sperm competition predicts continuous adaptation, where intensity is comparable with the degree of polyandry. Male competition can also drive adaptation of inseminated proteins, which affect female behavior, as it occurs in *Drosophila* accessory gland proteins. Inseminated proteins have been shown to affect the sperm storage in the female reproductive tract, copulatory plug formation, ovulation, oogenesis, female receptivity to re-mating, and female lifespan. These can be important effects for sperm competition and sexual conflict, both of which may drive adaptive evolution (Swanson, et al. 2001; Wolfner 2002; Swanson 2003b; Swanson, et al. 2004; Mueller, et al. 2005; Clark, Aagaard and

Swanson 2006). Similarly, in mammals genes encoding proteins involved in sperm-egg interaction and fertilization exhibit a particularly fast evolution. Some of them, such as *ZP3*, *ADAM1*, and *ADAM2*, have individually been shown to evolve under positive selection in humans, suggesting a role of positive selection on sperm-egg interaction (Swanson and Vacquier 2002; Swanson 2003a; Clark and Swanson 2005; Aagaard, et al. 2006).

Recently, various studies have been focused on describing the close relationship between immunity and reproduction. A striking link between these two functions lies in proteins and peptides with antimicrobial activities that are expressed in reproductive tissues of vertebrates and invertebrates. One of those regions that contain genes encoding for such proteins is located on the human chromosome 20q13 and it is named the *WFDC* locus.

1.2 The *WFDC* locus

1.2.1 Structure and Organization

The *WFDC* locus encodes for 17 small serine protease inhibitors that regulate endogenous proteases and is organized in two subloci, separated by 215 kb of unrelated sequence (Figure 1). The sublocus closer to the centromere (*WFDC*-CEN) includes *PI3* (also referred to as *elafin*) and *SLPI*, known for their anti-viral activity, and the two genes without any inhibitory activity, the semenogelins (*SEMG1* and *SEMG2*), which encode for the major semen proteins. The sublocus closer to the telomere (*WFDC*-TEL) includes the genes of *WFDC2* (also known as *HE4*), a recognized ovarian cancer marker, and *EPPIN* (also known as *SPINLW1*), the most studied serine protease inhibitor of the *WFDC*-TEL sublocus involved in male reproduction (Richardson, et al. 2001; Clauss, Lilja and Lundwall 2002; Hellström, et al. 2003).

This family appears to have a common origin through a series of duplications, but it gave rise to proteins with highly differing primary structure. The members of the *WFDC* locus have conserved first and last exons, preserving structures of importance for signaling, such as upstream promoter elements, signal peptide and 3' non-translated sequences. However, the second exon has undergone rapid evolution in such a way that there is almost no sequence similarity (Lundwall and Lazure 1995; Lundwall and Ulvsback 1996; Lundwall and Clauss 2011).

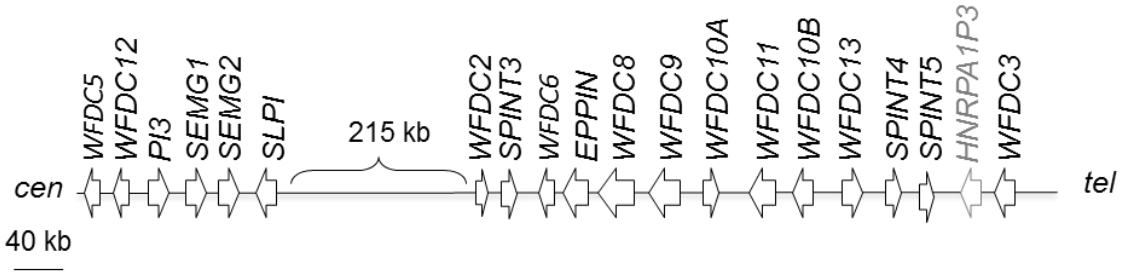


Figure 1: Schematic representation of the 20q13 *WFDC* locus. Diagram showing the relative positions of the *WFDC* genes, where the *WFDC* locus spans 700 kb. Its genes are organized into two subloci (centromeric and telomeric; *WFDC*-CEN and *WFDC*-TEL, respectively), separated by 215 kb of unrelated sequence. *HNRPA1P3* pseudogene is indicated in light gray.

Most *WFDC* genes are predominantly expressed in the male reproductive tract, even though some can be found in other tissues (Clauss, Lilja and Lundwall 2002). They exhibit functions involving reproduction, antimicrobial, immune, and tissue homoeostasis activities but, in many cases, WFDC biological functions remain poorly understood (Yenugu, et al. 2004; Bouchard, et al. 2006; Bingle and Vyakarnam 2008; Lundwall and Clauss 2011). These genes have in common the WAP domain, which was first identified in whey acidic protein, a predominant protein in the milk whey of lactating mice. It consists of ~40 amino acid residues with a characteristic disulfide pattern, which names the WFDC domains (Hennighausen, et al. 1982; Tsunemi, et al. 1996). They are organized in a central core β-sheet surrounded by two external peptide chains that are connected by a loop and hold the protease binding site (Figure 2).

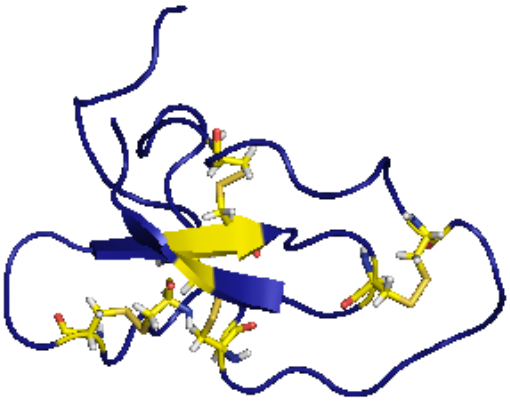


Figure 2: Three dimensional structure of the WAP domain from elafin (PDB reference: 2REL). β-sheets are represented in arrows and loops are represented by lines. The four disulfide bonds are highlighted in yellow.

Even though most of the *WFDC* genes have a WAP domain, some of them also contain additional domains, such as the Kunitz, trappin and semenogelin domains (Peter, et al. 1998; Schalkwijk, Wiedow and Hirose 1999; Clauss, Lilja and Lundwall 2002). Kunitz domains, which also confer serine protease inhibitor activity, are present in several *WFDC* genes, such as *WFDC6*, *EPPIN*, and *WFDC8*. In some cases, the WAP domain was lost and only the Kunitz domain prevails in the gene structure (*SPINT3*, *SPINT4* and *SPINT5*) (Petersen, et al. 1996; Clauss, Lilja and Lundwall 2002; Clauss, et al. 2011). Nonetheless, in the case of elafin, its secundary structure contains a trappin domain in addition to the WAP. The N-terminal region of elafin includes a variable number of repeats (Gly-Gln-Asp-Pro-Val-Lys) that act as an anchoring motif by transglutaminase crosslinking (Schalkwijk, Wiedow and Hirose 1999). *SEMG* genes share a common ancestor with the *WFDC*s, more specifically with elafin. They both belong to the *REST* family, where the first and last exons of all genes are highly conserved among mammals, while the second exon is typically highly diverged (Lundwall and Ulvsback 1996; Hurle, et al. 2007). The rapid evolution of the *SEMG* genes involves the expansion of the second exon, leading to a highly repetitive primary structure of the encoded SEMG proteins (Lundwall and Lazure 1995; Jensen-Seaman and Li 2003). These are composed of a 60 amino acid residue repeats, with high amino acid identity (78%), six in SEMG1 and eight in SEMG2 (de Lamirande, et al. 2001; de Lamirande 2007; Malm, et al. 2007).

1.2.2 Functions in reproductive biology

SEMGs and EPPIN have been shown to have important roles in fertility and sperm maturation. SEMG1 and SEMG2 are originated in the seminal vesicles, and are the most abundant proteins in the semen coagulum, contributing to approximately 60% of the ejaculate volume. The secretion from the epididymis that contains the spermatozoa, is only a small part of the ejaculate volume. The remaining fraction of the semen comes mainly from the prostate in a secretion rich in serine proteases and Zn^{2+}. At ejaculation the fluids are mixed and SEMGs form a gel, entrapping the spermatozoa. In humans, within 20 min of ejaculation, the gel is almost completely liquefied by serine proteases, mainly by PSA (also known KLK3), which cleaves SEMG1 and SEMG2 into small fragments. Simultaneously, the spermatozoa become more motile and matured, ready to fertilize the egg (Lilja, Abrahamsson and Lundwall 1989; Peter, et al. 1998; Robert and Gagnon 1999; de Lamirande, et al. 2001; Lundwall, et al. 2002; Jonsson, et al. 2006; de Lamirande 2007; Lundwall 2007; Jonsson, Frohm and Malm 2010). Several studies have shown that EPPIN protein is localized on the surface of ejaculated spermatozoa, in a complex of proteins containing LCF, CLU and SEMGs (Wang, et al. 2007b). This protein complex is capable of modulating PSA activity and provides antimicrobial protection for spermatozoa in the ejaculate coagulum (O'Rand, et al. 2006; O'Rand, et al. 2009; O'Rand, et al. 2011; Silva, et al. 2012). The importance of EPPIN in male reproduction was confirmed by contraceptive studies in nonhuman primates, which demonstrated that male rhesus monkeys immunized with recombinant human EPPIN were reversibly infertile (O'Rand, et al. 2004). A more recent study has shown that EPPIN residues Cys102, Tyr107, and Phe117, project for SEMG1 into a central binding pocket, are critical for EPPIN and SEMG1 interaction and could be used as potential targets for the design of contraceptive drugs (Silva, et al. 2012).

Interestingly, both EPPIN and SEMGs have an important role in fertility like maintaining the integrity of the spermatozoa and protecting them from microbial infections. When the SEMGs are cleaved by PSA, the resulting peptides from SEMG1 have been shown to have antimicrobial properties (Lundwall, et al. 2002; Bourgeon, et al. 2004; Edstrom, et al. 2008; Zhao, et al. 2008; Martellini, et al. 2009). Similarly, EPPIN has a demonstrated antibacterial activity in the reproductive tract (Yenugu, et al. 2004; Doumas, Kolokotronis and Stefanopoulos 2005; Wang, et al. 2005; McCrudden, et al. 2008). It is not surprising to find such role in EPPIN, as the WAP domains have

been shown to have antimicrobial functions not only in human proteins but also in other taxa (O'Rand, et al. 2011; Smith 2011; Wilkinson, et al. 2011).

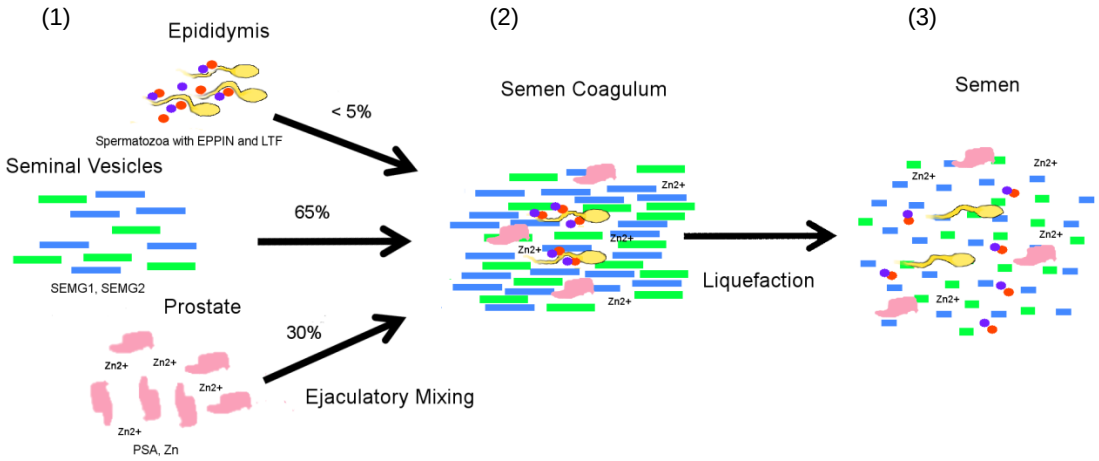


Figure 3: Schematic diagram showing the coagulation and liquefaction cascade in the human semen. **(1)** Different components of the semen are stored separately. Orange and Purple circles represent EPPIN and the LTF/CLU, respectively. They are mixed upon ejaculation. **(2)** The prostate secretion rich in Zn^{2+} and zinc-inhibited PSA is mixed with the seminal fluid. It contains large amounts of SEMGs that bind to the major fraction of Zn^{2+}. This induces a conformational change of SEMG1 enabling the formation of a gel, and decreases the concentration of free Zn^{2+}. PSA is activated with the decreased concentration of free Zn^{2+}. **(3)** PSA cleaves the SEMGs, which results in liquefaction of the gel. Motile spermatozoa and SEMG1 antimicrobial peptides from SEMG1 (green) are released (adapted from Malm et al 2007 and Lundwall and Brattsand 2008)

1.2.3 Functions in inflammation and immune response

In the *WFDC* locus, the cases of *elafin*, *SLPI*, and *WFDC2* clearly demonstrate the antimicrobial, antiviral and inflammatory roles of the WAP domains. Elafin and SLPI are pleiotropic molecules engaged in the surveillance against microbial infections at mucosal surfaces (Williams, et al. 2006; Weldon, et al. 2007; Weldon and Taggart 2007; Moreau, et al. 2008; McKiernan, McElvaney and Greene 2011). They have been consistently associated with anti-HIV activity and in the triggering of the inflammatory response (Williams, et al. 2006; Weldon, et al. 2007; Weldon and Taggart 2007; McKiernan, McElvaney and Greene 2011).

SLPI and elafin were first described as protease inhibitors while SLPI forms complexes with neutrophil elastase, chymotrypsin and trypsin, while elafin inhibits neutrophil elastase but has a more restricted spectrum of inhibition. They both

modulate chronic inflammation when binding to pathogen-derived molecules such as LPS and some GAGs (Wilkinson, et al. 2011). *SLPI* expression is enhanced in the presence of neutrophil elastase (Sallenave, et al. 1994), drugs and hormones, where the ability of glucocorticoids to induce SLPI may be partly responsible for their anti-inflammatory action. Even though SLPI and elafin seem to have similar functions, they respond differently to pro-inflammatory cytokines suggesting that they are not redundant molecules (King, Critchley and Kelly 2003; Williams, et al. 2006; Roghanian and Sallenave 2008; Sallenave 2010; Wilkinson, et al. 2011). SLPI and elafin also play an important role in the prevention of HIV transmission. SLPI interrupts the interaction of the HIV virus with receptors on the host cell (McNeely, et al. 1995; McNeely, et al. 1997; Wahl, et al. 1997) through its binding to receptors on the surface of human macrophages and by disrupting the binding of the HIV surface to these receptors (Ma, et al. 2004). On the other hand, elafin has been identified as an inhibitor of HIV infection in the female reproductive tract through direct interaction of elafin and HIV (Ghosh, et al. 2010; Drannik, Henrick and Rosenthal 2011).

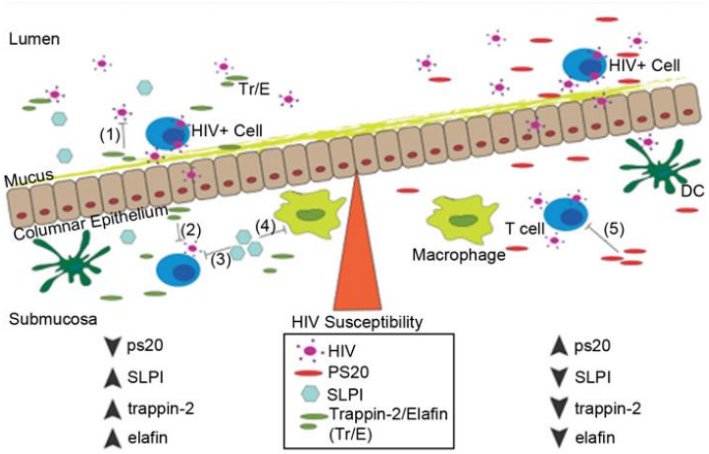


Figure 4: Potential model for Elafin/Trappin-2 (E/Tr) and SLPI modulation of HIV transmission, through viral recognition and antiviral immune-inflammatory response. **(1)** Tr and E inhibit HIV transcytosis through a monolayer of genital epithelial cells (ECs) by reducing attachment of virus to ECs; **(2)** E inhibits HIV infection of CD4+ T-cells by binding to HIV binding sites on T-cells; **(3)** SLPI inhibits HIV infection of macrophages by blocking binding between phosphatidylserine and annexin II; **(4)** SLPI inhibits HIV infection of CD4+ T-cells by preventing binding between HIV and CD4 responses (Drannik, Henrick and Rosenthal 2011).

Moreover, SLPI and elafin appear to have an important role in modulating immune response through the NF-κB signaling pathway. The activation of NF-κB leads to the transcription of numerous pro-inflammatory and cell-survival genes, however

when the NF-κB is upregulated it causes inflammatory and malignant diseases, such as breast, ovarian, prostate and lung cancer (Bouchard, et al. 2006; Cai, Tchou-Wong and Rom 2011). Concurrently, most types of ovarian cancer overexpress *SLPI* and non-small cell lung carcinoma patients have elevated levels of SLPI in the serum and in the tumoral tissues. On the other hand, *elafin* is upregulated in prostate-cancer cells when NF-κB signaling pathway is inhibited by genistein treatment, and both *elafin* and *SLPI* are downregulated in bladder cancer (Bouchard, et al. 2006).

Similarly, *WFDC2* is upregulated in ovarian cancer, is recognized to interact with inflammatory response proteins, such as cytokines and NF-κB, and is proposed as a biomarker for early stage tumors (Drapkin, et al. 2005; Bingle, et al. 2006; Bouchard, et al. 2006). Outside of the role in cancer progression WFDC2 has a large protease inhibitor spectrum and a wide activity against microbial virulence factors (Chhikara, et al. 2012) .

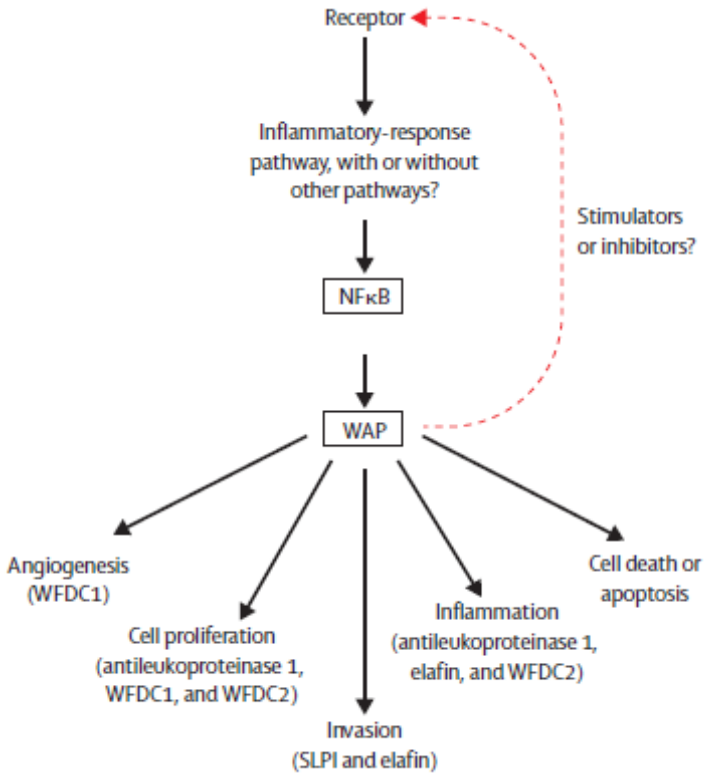


Figure 5: Potential WAP roles in cancer (Bouchard, et al. 2006).

Even though most of the *WFDC* genes are not fully functionally characterized, the high degree of domain conservation between them suggests a similar antibacterial activity and protease inhibitor functions.

1.3 *WFDC*s Comparative Genomics

The clustering of the *WFDC* genes in the human chromosome 20 suggests that they evolved by tandem duplications, where a possible early duplication created two genes that subsequently evolved into the *WFDC*-CEN and *WFDC*-TEL subloci. Such organization of the *WFDC* locus is present in all primates, rodent lineages, and several other mammals (Clauss, Lilja and Lundwall 2005; Lundwall and Clauss 2011). However, the low degree of sequence conservation between *WFDC* paralogs point toward an early origin in evolution (Clauss, Lilja and Lundwall 2002; Clauss, Lilja and Lundwall 2005).

The gene organization within the human *WFDC*-CEN is common to other mammalian lineages (Figure 6). In rodents, the functional equivalent of *SEMGs*, the seminal vesicle secreted proteins *Svs II-VI,* have expanded by gene duplication and in spite of lacking sequence similarity to *SEMGs*, their three dimensional structure is very similar. The *Svs* genes conserved location in the sublocus supporting the hypothesis that *SEMG*s and *Svs* evolved from a common *WFDC* ancestor gene (Clauss, Lilja and Lundwall 2005). The *WFDC*-CEN sublocus of rat and mouse are very similar but *Slpi* was detected to have four paralogs in rat, in which one copy carries the conserved *WFDC* domain with a high affinity for elastase, and the remaining encode inhibitors with affinities towards other proteases (Clauss, Lilja and Lundwall 2005). In addition to *Svs* and *Slpi* genes, the rodent *WFDC*-CEN includes the genes encoding SWAM1 and SWAM2 proteins that are orthologs to *WFDC15* pseudogene and *WFDC12*, respectively. Events of *WFDC* gain and loss are present in several other mammals. These include the example of *elafin* in the pig which has undergone at least three duplications and the case of *WFDC12* ortholog that was lost in the domestic dog. In humans, *WFDC12* is highly expressed in the prostate, and it would be expected to be conserved in dogs, however its role could have been replaced by the neighboring *Wfdc15*, which is pseudogenizated in humans. The domestic dog also lost *SEMG1* and *SEMG2* orthologs and has no functional equivalent to SEMGs probably due to the absence of seminal vesicles in this species (Clauss, Lilja and Lundwall 2005).

In the *WFDC*-TEL sublocus, three duplications can be traced through the levels nucleotide/protein conservation. A first, resulting in *EPPIN* and *WFDC6* genes and a

second leading to *WFDC9* and *WFDC11*, both in the same range of divergence from *SEMG1* and *Svs2*, suggesting that the duplication occurred around the time of the separation of the primate and murine lineages. The third duplication yield *WFDC10* and *WFDC12* genes and probably occurred before the separation of New World monkeys and catarrhine primates, as the conservation of nucleotides is of the same order as *SEMG1* genes from human and New World monkeys (Lundwall and Ulvsback 1996; Peter, et al. 1998).

In several instances, *WFDC* transcripts show large differences between species due to splicing. For example, mouse *Wfdc2* contains an extra exon adding a unique structure at the center of the molecule not present in other orthologs (Bingle, Singleton and Bingle 2002). In human *WFDC8*, the signal peptide originates from the first and the second exons, whereas in mouse *Wfdc8*, the first exon is a 5' UTR and the second exon is where the translation starts, and has no correspondence to human exons (Clauss, Lilja and Lundwall 2005).

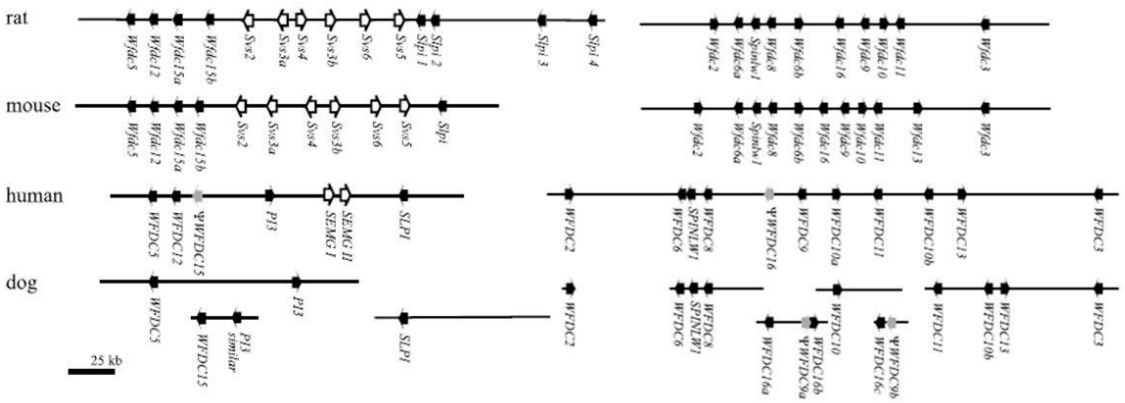


Figure 6: *WFDC* locus organization in mammals (adapted from Clauss et al. 2005). The arrows indicate the direction of transcription and approximate location of the genes. Black arrows indicate functional genes; grey arrows indicate pseudogenes and open arrows indicate genes enrolled in semen coagulum.

1.4 Primate Adaptive Evolution

Most of the primate evolution studies that focused on the *WFDC* locus were mainly centered in the *WFDC*-CEN, especially in *SEMG*s. These genes have been identified as genes under adaptive evolution, specifically by correlating their nucleotide sequences and copy-number variants to the different mating systems of various

primate species (Jensen-Seaman and Li 2003; Kingan, Tatar and Rand 2003; Dorus, et al. 2004; Carnahan and Jensen-Seaman 2008).

Specifically, in monoandrous primates like gorillas and gibbons, in which females mate with a single male, the ejaculate is gelatinous in texture and does not coagulate. Conversely, polyandrous primates (or multimale-multifemale), like chimpanzees and macaques, in which each female mates with multiple partners per ovulatory period, the ejaculate forms a strong and almost rigid copulatory plug that prevents the insemination of females by competing males (Dixson and Anderson 2002; Dixson and Anderson 2004). Several studies have shown that in chimpanzees the length of SEMG1 has radically expanded compared to other species, increasing the number of potential sites for transglutaminase crosslinking and protein–protein interactions, which are responsible for the formation of the copulatory plug (Jensen-Seaman and Li 2003). Interestingly, *SEMG1* has reduced levels of polymorphism in chimpanzees that were interpreted as an evidence of a selective sweep driven by the intense sperm competition in this species (Kingan, Tatar and Rand 2003). Other studies have shown that female promiscuity and semen coagulation in primates correlate with the rate of *SEMG2 and SEMG1* evolution (Dorus, et al. 2004; Hurle, et al. 2007). The pattern of strong sperm competition probably contributed to the rapid evolution of *SEMG*s (high d_N/d_S values) in more promiscuous primate species, while in the monoandrous species due to the presumed absence of sperm competition these proteins accumulated loss of function mutations (Hurle, et al. 2007; Carnahan and Jensen-Seaman 2008). Despite the controversial classification of the human mating system (monogamous or dispersed) SEMG1 and SEMG2 proteins were reported to have evolved under relaxed functional constraints in this species (Dixson and Anderson 2002; Dorus, et al. 2004; Hurle, et al. 2007). Other *WFDC*-CEN genes (*WFDC12*, *PI3*, and *SLPI*) were associated with high d_N/d_S ratios suggestive of strong patterns of adaptive evolution, even though an unambiguous correlation between gene mutation rates and mating systems was not confirmed (Hurle, et al. 2007). On the other hand, *WFDC*-TEL evolutionary history and interspecies comparisons has been only poorly explored, and to date no study has focused in the intraspecies variability of this sublocus (Clauss, Lilja and Lundwall 2005; Lundwall and Clauss 2011).

Overall, the organization of the *WFDC* locus suggests a gene evolution from a single ancestor by multiple duplications. As most of these genes are highly expressed

in the male reproductive tract, it is considered that gene duplications conserved the important regulatory elements. Post-copulatory selection has clearly shaped some of the genes in this locus, with the addition of several repeats in SEMGs. Furthermore, other selective pressures such as pathogen infections may have shaped the evolution of these genes. Thus future studies of *WFDC* genes should make available important insights into the molecular mechanisms behind their gene evolution.

2. Aims

There is a lack of information on the evolution of the *WFDC* locus in human populations and the forces that have shaped its notable genetic variability. Available studies are piecemeal and do not focus on the evolutionary scale of modern human populations, but rather on signatures of selection in a more distant timescale. In addition, the biological roles of most of the genes at the *WFDC* locus remain uncharacterized. Taking into account the implications of the *WFDC*s and *SEMG*s in human health, a better understanding of the evolutionary history of these genes and a way to prioritize which genes should be further assessed for their biological function is critical.

The study design included a high-throughput targeted sequencing of the *WFDC* locus, data from GWS scans for positive selection, and publically available genetic variation databases, which were combined and analyzed to unravel the adaptive forces shaping the evolution of this locus.

The aims of the project were:

1. Tease apart the signals of natural selection shaping the genetic diversity at the *WFDC* locus in human populations.

Studying the genetic diversity and patterns of selection in the *WFDC* locus was accomplished in two phases:

Stage I: A sequencing study to validate selective signals as proposed by a previous GWS for positive selection. This was a targeted, gene-centered approach, only performed in the populations where a selective signal was found.

Stage II: A high-throughput targeted sequencing study that systematically cataloged the genomic variation across the *WFDC* locus in three human populations. This included the sequencing of all the exons and several interspaced noncoding regions of our locus of interest and a set of control regions to determine the demographic model that better fit the data, and to detect potential outliers.

2. Characterize the extent of genetic diversity at the *WFDC* locus within and between hominids.

For this systematic comparative genomics and population-based genetics analysis the entire *WFDC* locus was sequenced in three chimpanzee subspecies (*P. t. troglodytes*,

P. t. verus and *P. t. ellioti*) and contrasted polymorphism and genetic variation at the human *WFDC* locus. This effort was aimed at learning whether the evolutionary forces driving the rapid diversification of *WFDC* and *SEMG* genes differ among hominid species.

3. Test the biological impact of the candidate variants on the genes under selection.

By undertaking molecular biology and expression assays, both *in vitro* and *in vivo* potential consequences of the selective signals found in hominids were sought.

It will increase the understanding of the complex evolutionary history of these genes.

3. Results

3.1 Evolutionary history of *WFDC*s in human populations

3.1.1. Differing evolutionary histories of *WFDC8* (short-term balancing) in Europeans and *SPINT4* (incomplete selective sweep) in Africans.

Mol Biol Evol. 2011 Oct;28(10):2811-22

Differing Evolutionary Histories of *WFDC8* (Short-Term Balancing) in Europeans and *SPINT4* (Incomplete Selective Sweep) in Africans

Zélia Ferreira,[1,2,3] Belen Hurle,[3] Jorge Rocha,[1,2] and Susana Seixas*,[1]

[1]Institute of Molecular Pathology and Immunology of the University of Porto, Porto, Portugal

[2]Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto, Portugal

[3]Genome Technology Branch, National Human Genome Research Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD

***Corresponding author**: E-mail: sseixas@ipatimup.pt.

Associate editor: Willie Swanson

Abstract

The whey acidic protein four-disulfide core (*WFDC*) gene cluster on human chromosome 20q13, harbors 15 small serine protease inhibitor genes with roles in innate immunity, reproduction, and regulation of endogenous proteases kallikreins. The *WFDC* cluster has emerged as a prime example of rapid diversification and adaptive evolution in primates. This study sought a better understanding of the evolutionary history of *WFDC* genes in humans and focused on exploring the adaptive selection signatures found in populations of European (Utah residents with ancestry from northern and western Europe [CEU]) and African (Yoruba from Ibadan, in Nigeria [YRI]) ancestry in a genome-wide scan for putative targets of recent adaptive selection. Our approach included resequencing coding and noncoding regions of *WFDC6*, *EPPIN*, and *WFDC8* in 20 CEU and of *SPINT4* in 20 YRI individuals. We generated 302 kb and 60 kb of high-quality sequence data from CEU and of YRI populations, respectively, enabling the identification of 72 single nucleotide polymorphisms. Using classic neutrality tests, empirical and haplotype-based analysis, we pinpointed *WFDC8* and *SPINT4* as the likely targets of short-term balancing selection in the CEU population, and recent positive selection (incomplete selective sweep) in the YRI population. Putative candidate variants targeted by selection include 44A (rs7273669A) for *WFDC8*, which may downregulate gene expression by abolishing the binding site of two transcription factors; and a haplotype configuration [Ser73+98A] (rs6017667A–rs6032474A) for *SPINT4*, which may simultaneously affect protein function and gene regulation. We propose that the evolution of *WFDC8* and *SPINT4* has been shaped by complex selective scenarios due to the interdependence of variant fitness and ecological variables.

Key words: *WFDC*, natural selection, innate immunity, serine protease inhibitor, reproduction.

Introduction

The availability of dense catalogues of human genomic variation has led to identification of intriguing "outlier" loci that present unusual patterns of genomic variation compared with the rest of the genome. These catalogues help unravel the confounding effects of natural selection (affecting genomic variation at specific loci) and population demographic history (shaping patterns of variation at all loci in a genome) (Biswas and Akey 2006; Biswas et al. 2009).

A growing number of genome-wide scans (GWS) for positive selection in humans suggest that certain types of genes are overrepresented among those that have been targets of positive selection: Among those are proteolysis genes and genes related to reproduction and immune functions (Wang et al. 2006; Akey 2009). Consistent with this, a recent GWS using the integrated haplotype score (iHS) and HapMap Phase II data (Voight et al. 2006) identified candidate genes bearing putative signals of recent positive selection within the whey acidic protein (WAP) four-disulfide core domain (*WFDC*) gene cluster on human chromosome 20q13 (fig. 1). Specifically, two signals of positive selection were found spanning the intervals from *WFDC6* to *EPPIN* (in populations of European descent or CEU) and from *WFDC10A* to *SPINT4* (in populations of African descent or YRI), respectively.

The *WFDC* cluster encompasses 15 genes encoding small serine protease inhibitors with characteristic WAP and/or for Kunitz domains that confer serine protease inhibitor and antibacterial activities (Clauss et al. 2005; Macedo-Ribeiro et al. 2008; McCrudden et al. 2008). Thus, a number of *WFDC* genes encode proteins with confirmed roles in innate immunity (*SLPI*, Zhu et al. 2002; *PI3*, Sallenave 2010), reproduction (*EPPIN*; O'Rand M et al. 2004), and regulation of the endogenous protease kallikreins (*KLK*, Lundwall et al. 2006). Although the functions of most other *WFDC* genes are poorly characterized, their similar structural domains suggest related functions. The neighboring seminal genes Semenogelin 1 and 2 (*SEMG1* and *SEMG2*) also play a central role in fertility and immunity (Lundwall 2007; Edstrom et al. 2008).

With their concomitant key roles in reproduction and innate immunity, *WFDC* and *SEMG* genes appear to occupy crossroads of multiple, interconnected biological processes

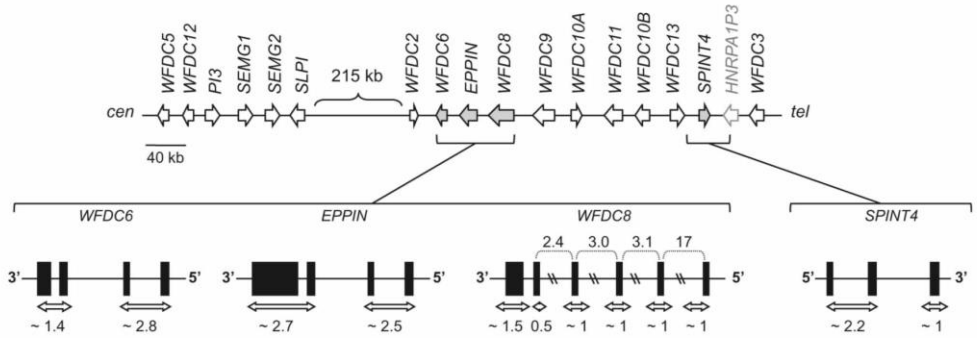


FIG. 1. Schematic representation of the *20q13 WFDC* gene cluster. Upper diagram shows the relative position of the *WFDC* genes. As depicted, the *WFDC* cluster spans ~700 kb and its genes are organized into two subclusters (centromeric and telomeric) divided by 215 kb of unrelated sequence. *WFDC* genes (including *SEMGs*) share conserved 5′ and 3′ UTRs, suggesting common origin from a single ancestral gene. The typical *WFDC* gene contains a promoter region, a 5′ exon coding for a signal peptide, one or more exons encoding WAP domains, and a 3′ exon with limited or no coding sequence. A *WFDC* gene can also include additional sequence coding for Kunitz structural domains (*WFDC6*, *EPPIN*, *WFDC8*, and *SPINT4*) or have lost their WAP domains all together (*SPINT4*). Gray arrows highlight surveyed *WFDC* genes. The insets show the exon/intron structure of each surveyed gene (exons are represented by solid boxes). Large arrows indicate the extension of the segments resequenced (size in kb). The localization of *HNRPA1P3* pseudogene is indicated in light gray.

including host-pathogen interactions, the effects of different mating systems on postcopulatory sperm competition, and male attempts to counteract the female immune response (Dorus et al. 2004; Hurle et al. 2007; Ramm et al. 2008). The relative contributions of these (or other) selective pressures to the overall evolution of these genes are currently unknown.

Complementary reports indicate that the number of genes under selection at the *WFDC* cluster may be quite large. For instance, initial comparison of the human and chimpanzee genome sequences identified the *WFDC* cluster as one of 16 genomic regions with an unusually high density of rapidly evolving genes (Chimpanzee Sequencing and Analysis Consortium 2005). Likewise, an independent study involving resequencing of the *WFDC* centromeric subcluster in 12 primates showed strong patterns of positive selection in a striking number of contiguous genes (*WFDC12*, *PI3*, *SEMG1*, *SEMG2*, and *SLPI*) (Hurle et al. 2007). Others have confirmed strong positive selection acting on *SEMG* genes (Kingan et al. 2003; Dorus et al. 2004; Ramm et al. 2008).

The present study sought a better understanding of the selection pressures acting on *WFDC* genes and focused on the positive selection signatures found in CEU and YRI populations by Voight et al. 2006. Our approach included reanalyzing HapMap Phase II haplotype data and resequencing the coding and noncoding regions of *WFDC6*, *EPPIN*, and *WFDC8* in 20 CEU samples and of *SPINT4* in 20 YRI samples. Individual |iHS| single nucleotide polymorphism (SNP) scores and classic neutrality tests enabled us to identify haplotypes linked to advantageous alleles that are consistent with positive selection centered in *WFDC8* and in *SPINT4*. We propose that *WFDC8* and *SPINT4* have been shaped by short-term balancing selection (in CEU) and by an incomplete selective sweep (in YRI), a likely result of the interdependence of variant fitness and ecological variables.

Our study is the first to identify *WFDC8* and *SPINT4* as bearers of unique, complex selective signatures that signal their prominent biological role among other *WFDC* genes, and the need for their further functional characterization. The results contribute significantly to mounting evidence that *WFDC*s are genes under strong adaptive pressures in primate evolution, diversifying in ways that are perceptible even within the short timescale of modern humans.

Methods

DNA Samples

All human samples used in the current study belong to the collection of the International HapMap Project Phase I/II. For the sequencing variation study, we surveyed a subset of 20 Europeans (CEU: Utah residents with ancestry from northern and western Europe) for *WFDC6*, *EPPIN*, *WFDC8* genes and another subset of 20 African individuals (YRI: Yoruba from Ibadan, in Nigeria) for *SPINT4* (supplementary table S1, Supplementary Material online). For typing the rs7273669 SNP, all available HapMap PhaseI/Phase II samples (232) were analyzed including those from Asians (Japanese from Tokyo, Japan [JPT]; and Han Chinese from Beijing, China [CHB]) (supplementary table S1, Supplementary Material online).

Polymerase Chain Reaction and Sequencing

Primers for amplification and sequencing were designed based on the GenBank (www.gdb.org) sequence entries for chromosome 20 (NC_000020.9) between bases 43596249 and 43787010. Sequencing was performed using the ABI BigDye Terminator version 3 cycle sequencing chemistry (Applied Biosystems, Foster City, CA), and electrophoretic analysis was done on an ABI 3130 automated sequencer. All human sequences were assembled and analyzed using the Phred-Phrap-Consed package (Nickerson et al. 1997). All putative polymorphisms and software-derived genotype calls were manually inspected and individually confirmed using

Consed. Further details of polymerase chain reaction (PCR) and sequencing conditions are available from the authors upon request.

Genotyping of the *WFDC8* Candidate Variant

The putative candidate variant found in the CEU samples (rs7273669A or -44A for brevity) was genotyped by PCR-restriction fragment length polymorphism analysis with the restriction enzyme *Bse*RI (New England BioLabs) in full HapMap Phase I/II sample collection (CEU, YRI, JPT, and CHB). The primers used for amplification of a fragment encompassing rs7273669 SNP were 5′-acatagaagttagagtggtc-3′ and 5′-acgcctaccttgctaaatgt-3′.

Statistical Analysis

Phased haplotypes from the International HapMap Project Phase II (release 21), for a 200 kb region centered on *EPPIN* in the CEU sample and on *SPINT4* in the YRI sample, were downloaded from the HapMap web site (http://hapmap.ncbi.nlm.nih.gov/). Haplotype data were then annotated with additional SNP information regarding ancestral allele state and potential selected sites. Ancestral allele state was retrieved from dbSNP (http://www.ncbi.nlm.nih.gov/), and potential selected sites were identified using the Haplotter application (http://hg-wen.uchicago.edu/selection/haplotter.htm), which displays the results of a GWS for positive selection based on the iHS statistic and relying on HapMap Phase I/II data sets (Voight et al. 2006). A $|iHS| > 2$ threshold, corresponding to the top 5% iHS values of the entire genome (Voight et al. 2006), was used to identify the potential selected sites.

Phased HapMap data for chromosome 20 were uploaded to Sweep 1.81 software (http://www.broad.mit.edu/mpg/sweep) and used to compute extended haplotype homozygosity (EHH) and relative extended haplotype homozygosity (REHH) (Sabeti et al. 2002). We defined cores as the longest nonoverlapping core haplotypes with at least three SNPs and no more than 20 SNPs, encompassing at least one SNP with significant iHS values, and containing the largest number of chromosomes carrying the candidate selected allele. Long-range haplotype tests were carried out at the largest physical distances, with REHH maximized and EHH close to 0.05. Significance of REHH, given the frequency of core haplotype, was calculated in Sweep 1.81 assuming 5% frequency bins.

For the *WFDC* resequencing survey, summary statistics of population genetic variation were calculated using SLIDER (http://genapps.uchicago.edu/slider/index.html); MAXDIP (http://genapps.uchicago.edu/labweb/index.html); and *H* test (http://www.genetics.wustl.edu/jflab/htest.html). Linkage disequilibrium (LD) and nucleotide diversity of the haplotypes were determined using the program DnaSP version 4.9 (Rozas 2009).

The haplotypes of the *WFD6-EPPIN-WFDC8* array and *SPINT4* were inferred by using the program PHASE 2.02 (Stephens et al. 2001; Stephens and Donnelly 2003), where SNPs previously inferred by the International HapMap Project Phase II were input as known phase to benefit from the trios genotyped in both CEU and YRI samples. To provide a temporal dimension to the phylogenetic relationships among haplotypes and to estimate the coalescent times and ages of relevant mutations, we used GENETREE version 9.0 (Griffiths and Tavare 1994). Five rare recombinant haplotypes carrying homoplasic mutations in the CEU population were removed from the *WFDC8* analysis. The mutation rate per gene, per generation, was deduced from the average number of nucleotide substitutions per site between human and chimpanzee reference sequences, calculated with DnaSP v.4.9 (Rozas 2009). Time estimates in generations were converted into years using a 25-year generation time. Human/chimpanzee divergence was assumed to have occurred 5.4 Ma (Patterson et al. 2006).

To assess the statistical significance of summary statistics, we ran 100,000 coalescent simulations (Hudson 2002) using estimates of the population recombination and mutation rate parameters calculated from our own data using MAXDIP and SLIDER. Simulations were produced using the "ms" program, assuming distinct demographic models including constant population size, expansion, bottleneck, structured population, and African and European best-fit models (Sabeti et al. 2002; Schaffner et al. 2005; Voight et al. 2005; Wang et al. 2006). For each model, we obtained a null distribution of summary statistics values and calculated the 2.5th and 97.5th percentiles.

Prediction of Transcript Factor Binding

Binding of transcript factors was predicted for both 5′ and 3′ untranslated regions (UTRs) of all the sequenced genes by the online available tool Mapper (Marinescu et al. 2005b). This program searches for putative transcription factor binding sites based on hidden Markov models built on alignments with experimentally known sites provided by the TRANSFAC (http://www.biobase-international.com/pages/index.php?id=transfac) and JASPAR (http://jaspar.cgb.ki.se/) databases.

Results

Selection Signatures at the *WFDC* Cluster According to HapMap Phase II Data

According to the GWS for recent positive selection based on the iHS statistic described by Voight et al. 2006 and HapMap Phase II data, two signals of positive selection reside in the *WFDC* cluster One spans genes *WFDC6* and *EPPIN* in CEU populations (significant empirical *P* values of $P = 0.0448$); another lies in the genomic segment containing *WFDC10A*, *WFDC11*, *WFDC10B*, *WFDC13*, and *SPINT4* in YRI populations (significant *P* values ranging from 0.0406 to 0.0441). For more detailed evaluation of the putative signals, we defined nonsliding windows of 200 kb centered either on *EPPIN* (CEU) or *SPINT4* (YRI) and mined phased HapMap Phase II data to identify all SNPs within the region with significant |iHS| values (supplementary files 1 and 2, Supplementary Material online). Consequently, ancestral allele state information was combined with the positive or negative significant |iHS| score of each SNP to determine

configurations of tightly linked alleles defining the longest haplotypes in the region. Hereafter, we will refer to extended haplotypes associated with a potential target of selection as "A" haplotypes and those unassociated as "B" haplotypes.

The extended "A" haplotype centered on *EPPIN* in the CEU population had an approximate frequency of 41% with very low genetic diversity (supplementary file 1 and supplementary fig. 1*A* and *B*, Supplementary Material online). Surprisingly, all 16 SNPs with significant |iHS| positive values within the 200 kb-long window were clustered with the *EPPIN* adjacent gene, *WFDC8* ($2.08 < |iHS| < 2.10$); in other words, none was linked to *WFDC6* or *EPPIN* as previous findings would have predicted (Voight et al. 2006). Thus, in subsequent studies, *WFDC8* was included as a putative target for recent positive selection in the CEU population.

The extended "A" haplotype centered on *SPINT4* in the YRI population was a short, ~30 kb fragment spanning *SPINT4* and nearby pseudogene *HNRPA1P3* (GI: 51511747) and included a large cluster of 30 SNPs with significant |iHS| scores ($2.0 < |iHS| < 3.6$) (supplementary file 2, Supplementary Material online). Given its frequency of approximately 80%, this "A" haplotype appears to present unexpectedly low variation (supplementary fig. 1*C* and *D*, Supplementary Material online). Thus, we hypothesize that *SPINT4* may be the single target of recent positive selection in the YRI population.

Resequencing for SNP Discovery at the *WFDC* Cluster

We next resequenced the *WFDC6-EPPIN-WFDC8* array and the *SPINT4* gene in a subset of 20 CEU and 20 YRI HapMap individuals, respectively. This step was necessary to avoid distortion of the allele frequency spectrum due to the ascertained nature of HapMap Phase II data and to uncover functional variation in the candidate genes. Figure 1 summarizes the resequencing strategy, which targeted coding exons and splicing junctions of the candidate genes but also included a number of small introns. Three tagger SNPs were also included in the resequencing survey to act as anchors for the "A" and "B" haplotype configurations as defined by HapMap Phase II data and to tag the two clusters of SNPs with significant |iHS| scores centered in either *WFDC8* (represented by tagger SNP rs6104221 in CEU) or *SPINT4* (represented by tagger SNPs rs1386504 and rs6032474 in YRI), respectively.

Overall, we generated 302 kb of high-quality sequence data from the CEU sample and 60 kb from the YRI sample, enabling identification of 72 SNPs (fig. 2). Importantly, 24 (33%) of those 72 sites lacked any previously associated reference SNP identification code in dbSNP (http://www.ncbi.nlm.nih.gov/projects/SNP/); and 19 (26%) of them, although tagged with a dbSNP number, had never been typed by the HapMap initiative. Upon integrating the new SNPs into the HapMap Phase II shell, we reexamined the haplotypes and nonascertained patterns of diversity across the *WFDC6-EPPIN-WFDC8* array and the *SPINT4* gene. Specific findings for *WFDC8* (in the CEU population) and *SPINT4* (in the YRI population) are discussed below.

WFDC8 in the CEU Population

Our sequencing and genotyping initiatives more than doubled the SNPs typed across the *WFDC6-EPPIN-WFDC8* gene array, exposing a total of 18 SNPs that differentiate the "A" and the "B" haplotypes. Initially, two positions in complete LD with the "A" haplotype stood out for their potential functional consequences: nonsynonymous amino acid replacement Asn137Ser in *WFDC8*; and SNP rs7273669G/A located 44 bp upstream from the transcription-start site of *WFDC8* (hereafter −44A for brevity).

However, upon closer examination, Asn137Ser was considered an unlikely target of selection for several reasons: The allele associated with the "A" haplotype is the ancestral one; codon 137 is not conserved among *WFDC8* orthologues; and the amino acid replacement is not predicted to affect protein stability or structure, as determined by Polyphen analysis (Ramensky et al. 2002).

On the other hand, −44A was predicted by Mapper (Marinescu et al. 2005a) to abolish, in the selected "A" haplotype, the binding sites of two ubiquitously expressed transcript factors—the myeloid zinc finger (MZF) and the peroxisome proliferator activated receptor (PPAR). Interestingly, G to A substitutions in transcription factor binding sites strongly contribute to loss of function and are more prone than other substitutions to alter gene expression (Lapidot et al. 2008). No functional variants in LD were detected with the "A" haplotype in association with *WFDC6* and *EPPIN*. Thus, we regard −44A as the best candidate variant among those that have been identified for being targeted by selection at the *WFDC6-EPPIN-WFDC8* gene array. The status of −44A as putative candidate variant warranted extended genotyping of this SNP in 232 HapMap Phase I/II individuals (see supplementary table S1, Supplementary Material online). When all available −44G/A data were inserted into the HapMap Phase II haplotypes, the −44A allele was not found in the CHB and JPT populations but was present in the YRI and the CEU populations (15% and 41%, respectively).

We questioned whether the long-range haplotypes would be maintained when haplotype analysis was centered on −44A. Thus, the haplotype structure across the *WFDC8* locus was reanalyzed with the EHH and REHH statistical tests, which use multiple SNPs in a core to increase statistical power (Sabeti et al. 2002; Sabeti et al. 2005) (fig. 3). In the CEU population, three major core haplotypes were identified, spanning positions from −4672 to +8722 (SNPs rs6032336 to rs6104239) relative to the ATG start codon of *WFDC8*. The observed frequencies reached 39% for the haplotype associated with the derived allele −44A (core GTTAGTTGAAGCAGAAGTTC), and frequencies of 29% and 14% for the two haplotypes associated with the ancestral allele −44G (cores GCTAATCAAGGCGGATATTC and GCTAAACAAGGCGGATATTC, respectively; fig. 3*A*). Moreover, the high REHH values of core GTTAGTTGAAGCAGAAGTTC extend an additional 410 kb (or 0.15 cM) on the telomeric side of *WFDC8*, with a value of 5.096—significantly elevated when compared with the distribution of REHH scores for HapMap Phase II data from

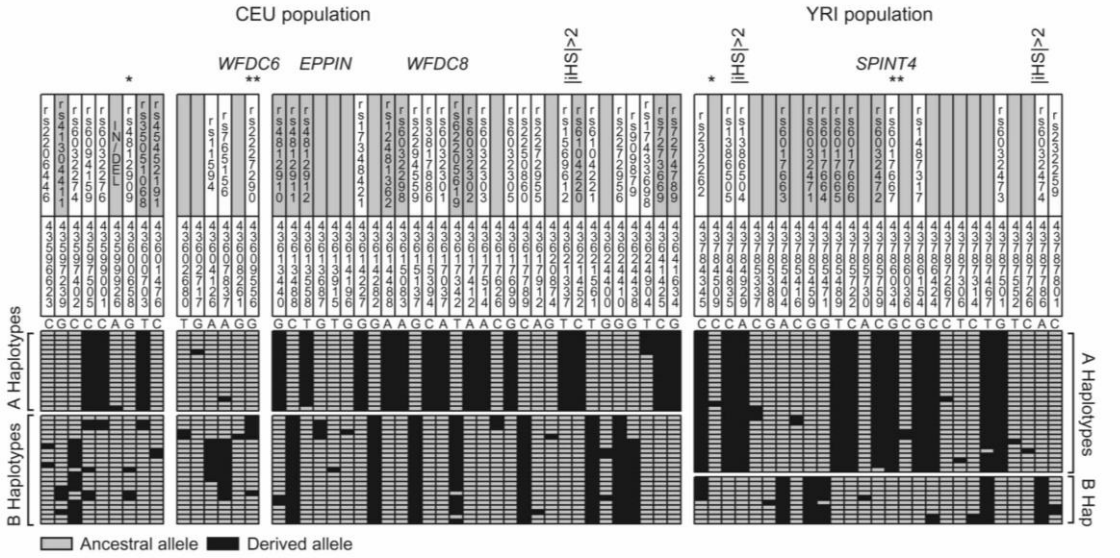


FIG. 2. "A" and "B" haplotypes for *WFDC6*, *EPPIN*, *WFDC8*, and *SPINT4*,as inferred by PHASE2.02 software. The ancestral state at each site was inferred based on orthologue nonhuman primate sequences. SNPs with significant |iHS| statistics are identified. Coding variants are marked by an asterisk. These included, in the CEU population, one synonymous amino acid replacement in *EPPIN* (Thr10Thr) and two nonsynonymous amino acid replacements in *WFDC8* (Asn137Ser and Thr96Met); and in the YRI population three nonsynonymous replacements in *SPINT4* (Ala30Glu, Gly73Ser, and Ser73Arg). SNP identifiers and their chromosomal positions based on NC000020 reference sequence are indicated in columns. SNPs typed in HapMap Phase II are in a white background. SNPs not typed by HapMap with a dbSNP reference number are in light gray background.

chromosome 20 and for the haplotype cores found at equivalent frequencies (35–40%; $P = 0.034$; fig. 3*B*). A similar analysis in the YRI population, centered on position −44, found narrow cores, short haplotypes, and nonsignificant REHH *P* values (result not shown).

Using classic neutrality tests and empirical distribution analyses, *WFDC8* significantly departed from neutral expectations in the CEU population. For instance, *WFDC8* showed unusually high nucleotide diversity ($\pi = 21.0$) and a strong positive value (2.38) of Tajima's *D* (Tajima 1989). Meanwhile, flanking genes *WFDC6* and *EPPIN* yielded unremarkable values of π and Tajima's *D* (table 1). Also for *WFDC8*, both statistics significantly deviated from neutrality in seven out of eight theoretical null distributions generated by coalescent simulations that assumed distinct demographic models (table 2). The departure from neutrality was particularly significant when the demographic model by (Schaffner et al. 2005) was tested with parameters calibrated to better fit the CEU data from HapMap Phase I. Furthermore, an empirical comparison with the subset of 316 genes from the SeattleSNPs project (http://pga.gs.washington.edu/; Crawford et al. 2005) that have been sequenced in CEU samples confirmed the outlier nature of *WFDC8*, with π and Tajima's *D* values falling within the 97.5th percentiles of the two empirical distributions in SeattleSNPs (supplementary fig. 2*A* and *B*, Supplementary Material online). Thus, we concluded that *WFDC8* significantly deviates from global trends of diversity in Europeans, with π and Tajima's *D* values indicating high sequence divergence and excess intermediate frequency alleles in the region.

We reconstructed the gene genealogy of *WFDC8* and estimated the time to most recent common ancestor (T_{MRCA}) using a maximum likelihood coalescent analysis (Griffiths and Tavare 1994). Interestingly, *WFDC8* showed an atypical structure dominated by two deep-root branches with intermediate frequencies. Each branch precisely segregated with either the potential selected allele −44A or the opposite allele −44G (fig. 4). Significantly, the tree topologies of adjacent genes in the same haplotype block *WFDC6* and *EPPIN* (supplementary fig. 3, Supplementary Material online) were in conformity with collected statistics for genes evolving under neutral evolution (Excoffier 2002; Tishkoff and Verrelli 2003; Satta and Takahata 2004; Garrigan and Hammer 2006). Given that previous findings (significant values of iHS, REHH, Tajima's *D*, and π), support a nonneutral interpretation of the evolution of *WFDC8*, the uncommon tree structure can be attributed to balancing selection acting exclusively on *WFDC8*. The recent T_{MRCA} estimate of 1.40 ± 0.27 Ma, coupled with the atypical deep-root tree topology, does not suggest the effect of ancestral population subdivision, as genes with deep-root branches affected by population structure in humans have longer T_{MRCA} estimates of 2–3 Ma (Hey and Harris 1999; Garrigan and Hammer 2006; Hayakawa et al. 2006; Patin et al. 2006; Kim et al. 2010).

SPINT4 in the YRI Population

The density of SNPs typed across the *SPINT4* locus increased from eight SNPs typed in HapMap Phase II to a total of 27 SNPs. The divergence of the "A" and "B"

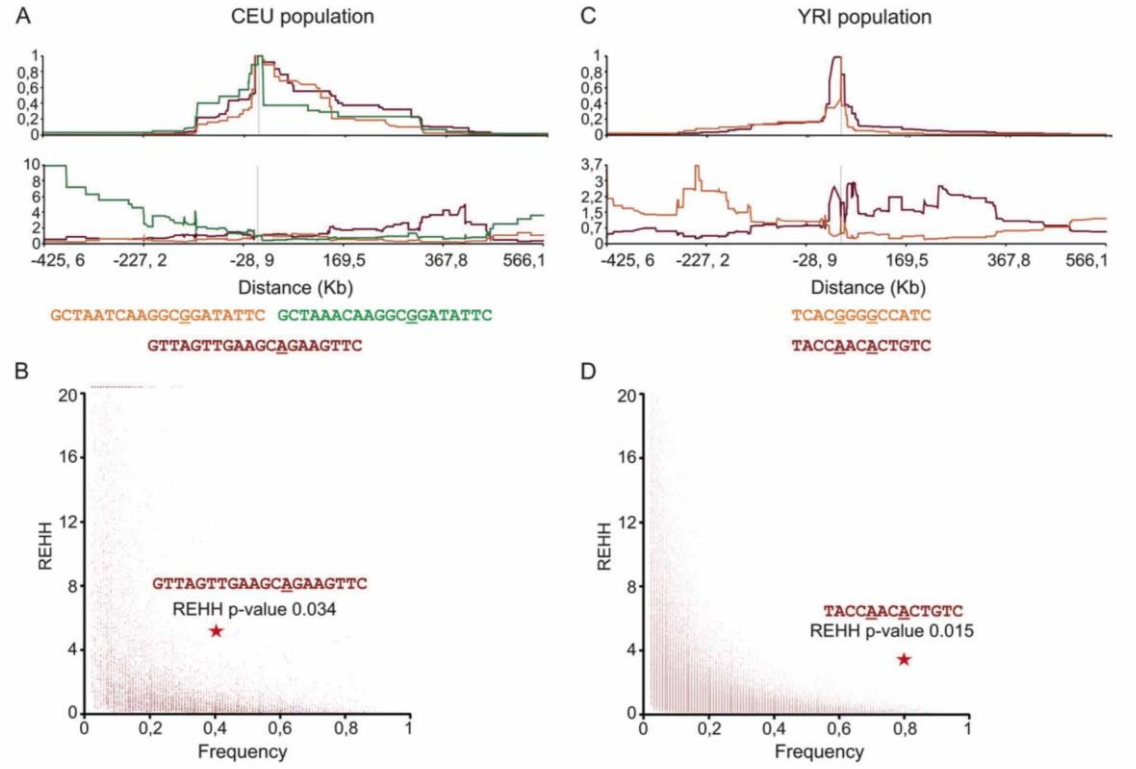


FIG. 3. Haplotype-based tests of selection using HapMap Phase II data. Plots of EHH and REHH over physical distance for the largest nonoverlapping cores encompassing the −44G/A (rs7274789) *WFDC8* variants (core haplotypes from rs6032336 to rs6104239 SNPs) (*A*) and for Gly73Ser+98A/G *SPINT4* variants (rs11908541–rs6130908) (*C*). Plots of REHH versus frequency for chromosome 20 using the CEU sample and a 410 kb distance (***B***) and using the YRI sample and a 150 kb distance (*D*). Core haplotype sequences are indicated below REHH plots and candidate variants are underlined. The stars represent the results for −44A bearing chromosomes (*B*) and for [Ser73+98G] bearing chromosomes (*D*).

haplotypes was determined by a total of 11 SNPs, two of which had potential functional consequences for *SPINT4*: a nonconservative change at position Gly73Ser and noncoding SNP rs6032474A/G in the 3′UTR. Conveniently, both candidate SNPs had been surveyed by the HapMap Phase II project in all HapMap individuals. The derived allele Ser73 introduces a nonconservative change at a highly conserved codon within the Kunitz domain, which in turn may significantly affect *SPINT4* structure and stability. Likewise, derived allele +98G is predicted to abolish a binding site for the Forkhead Box J2 (FOXJ2) transactivator factor in the 3′ UTR of *SPINT4*. Interestingly, in the YRI population, the derived allele Ser73 is in complete LD with the ancestral allele +98A ($|D'| = 1$, $r^2 = 1$), making it infeasible to dissociate the selective advantages of each individual allele. Thus, the best candidate for selection in the *SPINT4* locus may be a two-SNP haplotype configuration rather than a single allele. For brevity, we will refer to the proposed advantageous variant as [Ser73+98A] and to the opposite variant as [Gly73+98G]. The [Ser73+98A] signature is prevalent in the YRI population, at 80% frequency, significantly higher than its modest frequency in CEU (36%) and JPT+CHB (33%) populations.

EHH/REHH statistics identified two major haplotype cores in the YRI population. These spanned positions +109 to +4776 (or SNPs rs11908541–rs6130908) with respect to the start codon of *SPINT4*, and included 13 SNPs (three with significant iHS values). The observed frequencies reached

Table 1. Summary Statistics of Population Variation.

Population	Gene	N^a	L^b	S^c	$\theta_w{}^d$	π^e	D^f	H^g	ρ^h
Europeans (CEU: Utah residents with ancestry from northern and western Europe)									
	WFDC6	40	4243	9	4.98	6.08	0.65	0.41	1.41
	EPPIN	40	5288	6	2.67	2.43	−0.24	0.69	6.61
	WFDC8	40	5675	30	12.4	21.0	2.38*	0.35	0.20
Africans (YRI: Yoruba from Ibadan in Nigeria)									
	SPINT4	40	2877	27	22.06	21.19	−0.134	−1.99*	1.81

[a] N = number of chromosomes.
[b] L = total number of sites surveyed.
[c] S = number of segregating sites.
[d] Watterson's estimator of θ ($4N_e\mu$) (Watterson 1975) per base pair ($\times 10^4$).
[e] Nucleotide diversity per base pair ($\times 10^4$).
[f] Tajima's D statistic (Tajima 1989).
[g] Fay and Wu H statistic (Fay and Wu 2000; Zeng et al. 2006).
[h] Hudson's estimator of ρ ($4N_e r$) per base pair ($\times 10^4$) based on a conversion-to-crossover ratio of 2 and a mean conversion tract length of 500 bp (Frisse *et al.* 2001; Hudson 2001).
*Statistically significant $P < 0.05$.

Table 2. Percentile 97.5 of the Null Distributions Generated by Coalescent Simulations.

Model	*WFDC8* (CEU)		*SPINT4* (YRI)
	D [a]	π [b]	π [b]
Constant	1.78*	10.70*	9.48*
Recent expansion[c]			
(N_0= 10^4, N_1= 10^7, and t= 1,000 g)	0.91*	8.93*	8.83*
2-fold growth[d]			
(N_0 ~10^4 and t=1000 g)	1.74*	10.43*	9.32*
Short and severe bottleneck[d]			
(N_0 ~10^4; t_0=1,600 g; b= 0.1 t_1= 1,200 g)	2.25*	11.67*	10.43*
Long and mild bottleneck[d]			
(N_0 ~10^4; t_0=1,600 g; b= 0.4 t_1= 1,200 g)	1.11*	9.35*	8.34*
Structure[c]			
(*npop* =2 and *m* =1.0)	2.08*	11.31*	9.91*
(*npop* =3 and *m* =1.0)	2.15*	11.45*	9.95*
Structure[e]			
(*npop* =2 and *m* =0.5)	2.43	12.02*	10.60*
Best fit[f]	1.93*	11.03*	8.44*

NOTES.—*N*—effective size; *t*—time in generations; *b*—bottleneck intensity; *npop*—number of populations; *m*—migration rate per generation.
[a] Tajima's *D* statistic (Tajima 1989).
[b] Nucleotide diversity per base pair (×10^4).
[c] Model from Wang et al. 2006.
[d] Model from Voight et al. 2005.
[e] Model from Sabeti et al. 2002.
[f] Model from Schaffner et al. 2005.
*Statistically significant—the observed statistic is higher than the 97.5th values.

77% for core TACCAACACTGTC, including the proposed advantageous variant [Ser73+98A], and 15% for core TCACGGGGCCATC, including the opposite variant (Gly73+98G). Furthermore, core TACCAACACTGTC extended an additional 150 kb (0.02 cM) on the 3′ side of *SPINT4* with stronger LD and higher EHH/REHH statistics (REHH = 2.24) than core TCACGGGGCCATC (Fig. 3*C*). This REHH value was also significantly higher than the REHH values for other haplotype cores found at equivalent frequencies on chromosome 20 (75–80%; P = 0.015; Fig. 3*D*). By contrast, in non-African populations, core TACCAACACTGTC did not yield significant REHH *P* values.

SPINT4 deviated from neutral expectations in all eight theoretical null distributions of π generated by coalescent simulations assuming distinct demographic models (table 2). The *H* test (Fay and Wu 2000; Zeng et al. 2006) was also significant for *SPINT4* (−1.99; P = 0.032). Unexpectedly, when we estimated the Tajima's *D* distribution for the YRI population, *SPINT4* showed unremarkable, slightly negative Tajima's *D* values that generally followed the trend in the YRI population (supplementary fig. 2*C*, Supplementary Material online); however, when individuals were sorted with respect to intrahaplotype diversity, the haplotype linked to [Ser73+98A] was associated with significantly negative Tajima's *D* and *D** statistics of Fu and Li (Fu and Li 1993), whereas the haplotype linked to [Gly73+98G] was not (table 3). Importantly, the haplotype linked to [Ser73+98A] was more skewed toward rare variants (six singletons and two doubletons) and lower nucleotide diversity than was the opposite haplotype (supplementary fig. 4, Supplementary Material online). Empirical comparisons of *SPINT4* with the subset of 316 SeattleSNPs genes studied in the YRI population (103 genes) or African-American population (213 genes) showed that *SPINT4* had an unusually high π value achieved by few other SeattleSNPs genes (supplementary fig. 2, Supplementary Material online). In conclusion, *SPINT4* presents a significant *H* test, a high π value deviating from the global trends of diversity in Africans, and a haplotype linked to [Ser73+98A] that shows strongly negative Tajima's *D* and *D** values of Fu and Li, hinting at a nonneutral haplotype configuration.

As in the case of *WFDC8*, the tree topology of *SPINT4* was dominated by an atypical pattern with two deep-root branches (fig. 4). However, in the case of *SPINT4*, the higher T_{MRCA} estimate of 3.07 ± 0.55 Ma could theoretically be attributed to either selection or ancestral population substructure. Likewise, the branch associated with the candidate signature [Ser73+98A] was linked to a "star"-shape genealogy, which might also be connected with either a positive selective event (occurring about 0.93 ± 0.34 Ma for node A and 0.58 ± 0.18 Ma for node B) or with population expansion. When a selection parameter β (Coop and Griffiths 2004) is taken into account in reconstructing the coalescent process, positive selection is more likely than no selection. In addition, when selection is assumed, minimal age estimates of the candidate signature [Ser73+98A] are significantly reduced, bringing the selective event to a more recent time (~0.26 ± 0.05 Ma, supplementary fig. 5, Supplementary Material online). Also taken into consideration was that Africa's demographic history does not suggest that population substructure could significantly account for two highly divergent haplotypes segregating at intermediate frequencies in the YRI population. Taken together, the tree topology and unusual statistics suggest that *SPINT4* is undergoing an incomplete selective sweep.

Discussion

GWS have made a critical contribution to understanding the genetic bases of natural selection (Biswas and Akey 2006; Scheinfeldt et al. 2009). Nevertheless, it is essential to validate the putative selective signals and follow up with detailed case-by-case gene analysis (Biswas and Akey 2006; Teshima et al. 2006). The *WFDC* cluster, which harbors genes involved in reproduction and immunity, was identified as containing putative targets of recent positive selection in a GWS based on iHS statistics. In this study, we pinpoint *WFDC8* and *SPINT4* as the likely targets of balancing selection in the CEU population, and recent positive selection in the YRI population, respectively. In addition, we identify putative allele variants with potential functional consequences that might confer selective advantage.

Not surprisingly, iHS statistics performed on individual SNPs prevailed over gene-centered (window-based) iHS statistics to narrow down the selective signals from candidate regions to individual genes. This is most likely due to ascertainment bias and factors such as gene size and gene

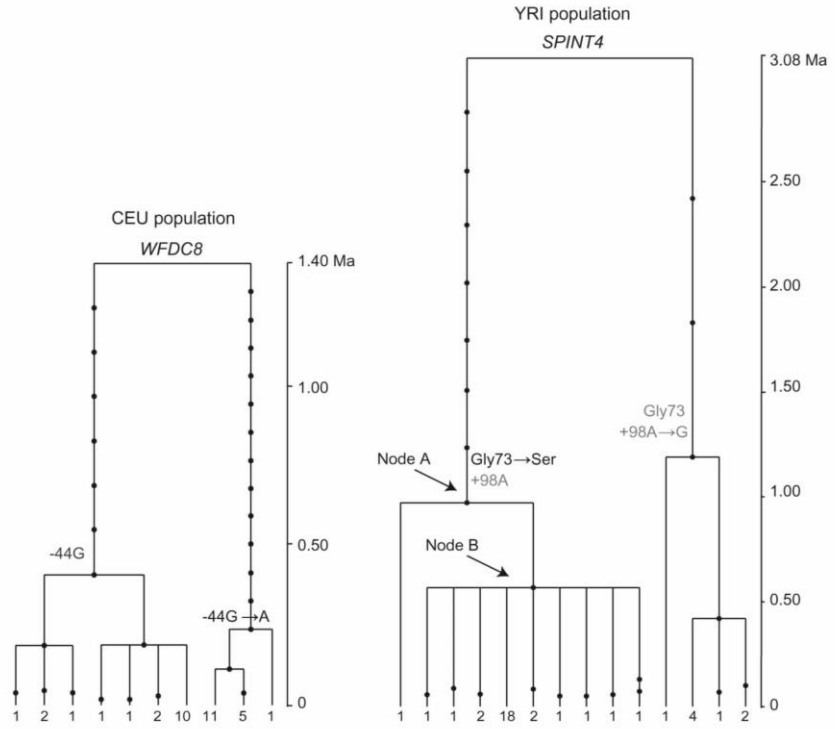


FIG. 4. *WFDC8* and *SPINT4* gene genealogies as estimated by Genetree. Time is scaled in millions of years (Ma). Tree branches corresponding to *WFDC8* −44G/A and to *SPINT4* Gly73Ser +98G/A variants are indicated.

density that influence or distort iHS statistics on window-based analyses (Seixas et al. 2007). For instance, gene-centered iHS analysis identified *WFDC6* and *EPPIN* as likely targets of selection in the CEU population; yet iHS statistics on individual SNPs indicated nearby *WFDC8* (with lower iHS values) as the likeliest target for selection. Similarly, it was not *WFDC11* or *WFDC10B* within the corresponding window, but rather *SPINT4* that had a more significant *P* value and the strongest evidence of selection in the YRI population.

Our results indicate that the −44A *WFDC8* allele (perhaps affecting the *WFDC8* regulatory control) has reached intermediate frequencies of ~40% in Europe; and the haplotype configuration [Ser73+98A] (presumably modifying *SPINT4* function and regulation) is nearing fixation in Africa, with frequencies of ~80%. However, the sequence variation and haplotype patterns suggest complex selective histories for *WFDC8* and *SPINT4* that are difficult to resolve by modeling standard scenarios of recent positive selection or long-term balancing selection.

It is possible that the selection signals are difficult to tease apart because *WFDC8* and *SPINT4* are evolving under complex adaptive scenarios for which statistics that rely on the frequency spectrum of sequence variation have limited sensitivity. Several alternative selective scenarios could explain the large π values associated with *WFDC8* and *SPINT4*. For example, in an ongoing selective sweep—with a new, rare allele being rapidly selected from introduction to fixation—π might increase its value as the allele reaches intermediate frequencies (Fay and Wu 2000; Przeworski et al. 2005; Zeng et al. 2006; Gordo et al. 2009). Alternatively, directional selection on standing variation—in which a neutral or mildly deleterious allele arises and drifts in the population, becomes beneficial at frequency *f*, and eventually reaches fixation—can be associated with large π values (Przeworski et al. 2005; Chevin and Hospital 2008). Thus, depending on the value of *f*, it can be challenging to discriminate between signals that indicate selective sweep acting on a newly arisen allele or selection on standing variation on a "rare" allele (Przeworski et al. 2005). It is also possible that incomplete sweeps and initial stages of balancing selection might produce quite

Table 3. *SPINT4* Intrahaplotype Diversity.

Haplotype	N^a	S^b	$\theta_w{}^c$	Π^d	D^e	D^*
***SPINT4* (YRI sample)**						
A	**30**	**10**	**8.77**	**3.16**	**−2.02***	**−2.36***
B	**10**	**7**	**8.60**	**7.72**	**−0.43**	**−0.13**

[a] N = number of chromosomes.
[b] S = number of segregating sites.
[c] Watterson's estimator of θ ($4N_e\mu$) (Watterson 1975) per base pair ($\times 10^4$).
[d] Nucleotide diversity per base pair ($\times 10^4$).
[e] Tajima's *D* statistic (Tajima 1989).
[f] Fu and Li (Fu and Li 1993).
* Statistically significant $P < 0.05$.

similar overall frequency spectrum patterns (Fay and Wu 2000; Przeworski et al. 2005; Charlesworth 2006; Zeng et al. 2006; Gordo et al. 2009)

For *WFDC8*, the remarkable divergence between "A" and "B" haplotypes at first resembled a prototypical pattern of long-term balancing selection. However, the homogeneity and length of the "A" haplotype spanning *WFDC8* was unusual: unlikely to result from local low recombination or even to derive from the ancestral structure of the CEU population (Schaffner et al. 2005; Garrigan and Hammer 2006; Gutenkunst et al. 2009). Instead, we believe that the extended haplotype observed for *WFDC8* is likely to correlate with the effects of "short-term" balancing selection. Several arguments support our hypothesis: The haplotype extension is not spurious, but allele-specific and linked to −44A; the selective signal is clearly captured through iHS or REHH statistics for which variation in local recombination is controlled; −44A has few linked polymorphic sites suggesting a recent, nonneutral increase in frequency; and the recent T_{MRCA} estimate of 1.40 ± 0.27 Ma, coupled with the atypical deep-root tree topology, does not suggest the effect of ancestral population subdivision (>2 Ma) (Hey and Harris 1999; Garrigan and Hammer 2006; Hayakawa et al. 2006; Patin et al. 2006; Kim et al. 2010). Similarly, the summary statistics π and Tajima's *D* are not in full concordance with the expectations of recent positive selection favoring a newly arisen variant. Considering all these observations, we propose a short-term balancing selection event for *WFDC8* with the following characteristics: an "A haplotype" swept to intermediate frequencies; linked variation including both rare and common variants (as if the selective sweep was recently stabilized); linked neutral common variation starting to accumulate in the newest allele; and differences between alleles not yet eroded by recombination.

SPINT4, on the other hand, presented unusually high values of π, an overall slightly negative Tajima's *D* value, and a significant *H* test value, all of which may be found when an advantageous allele is increasing in frequency toward fixation (Fay and Wu 2000; Przeworski et al. 2005; Zeng et al. 2006; Gordo et al. 2009). Moreover, all tests centered on "A" haplotype aspects, including EHH/REHH, Tajima's *D*, and *D** of Fu and Li, support *SPINT4* as a gene under an incomplete selective sweep—a new allele increasing its frequency rapidly toward fixation, thus becoming associated with an excess of rare variants. Interpretation of the selective scenario of *SPINT4* could be further complicated by the existence of two candidate variants in complete LD, or the haplotype configuration [Ser73+98A]. Still, these were traced to a single deep-rooted branch in the *SPINT4* tree, and to a "star"-shape genealogy typical of a positive selection event. Moreover, the literature describes other cases in which a two-SNP haplotype configuration has been swept to higher frequencies by natural selection. One example is a two-SNP haplotype configuration [Val72-Ile93] in the *N*-acylsphingosine amidohydrolase (*ASAH1*) gene involved in hydrolysis of ceramides and regulation of neuronal development (Kim and Satta 2008).

Regardless of the complexity of the underlying selective scenario(s), a selective (nonneutral) interpretation of the highly divergent lineages observed at the *WFDC8* and *SPINT4* loci is more plausible than alternative models based solely on neutrality or ancestral population substructure. The signals of selection on *WFDC8* and *SPINT4* were initially identified through a GWS for positive selection in which possible effects of demography, recombination, and SNP density and the allele frequency spectrum were somewhat controlled, making a selective interpretation more plausible than demographic effects. In comparisons with null distributions generated by neutral coalescent simulations under different demographic scenarios, and in empirical comparisons with SeattleSNPs genes, both genes are almost invariably "outliers," which also supports a selective interpretation. Finally, the tree topologies of two genes within the same haplotype block as *WFDC8* (*WFDC6* and *EPPIN*) conformed with collected statistics for genes evolving under neutral evolution—making the very restricted, locus-specific signal on *WFDC8* harder to explain under neutrality or demographic hypotheses alone.

Expression of *WFDC8* and *SPINT4* is mainly restricted to testis and epididymis, with the gene activity potentially extending over a large area in the male reproductive tract. As with other serine protease inhibitors (i.e., *SERPINA5*, *SERPINE2*, *EPPIN*, *SLPI*, and *PI3*, Uhrin et al. 2000; Murer et al. 2001; O'Rand et al. 2007; Wang et al. 2007a, 2007b; McCrudden et al. 2008), *WFDC8* and *SPINT4* are likely to play a role in regulating proteolysis cascades linked to the maturation and capacitation of sperm cells or they may be related to innate immune function in the male genital tract. For instance, *WFDC8* expression in epididymis cauda is lost after vasectomy and may contribute to the impaired fertility after vas deferens reanastomosis (Thimon et al. 2008); and a GWS identified the Gly73Ser (rs6017667) allele of *SPINT4* as being associated with the multifactorial autoimmune disease, Type I diabetes (T1D) (Todd et al. 2007). Interestingly, T1D has also been associated with impairments of male reproductive function in humans and in animal models, which show structural modifications of testis and epididymis and damaged sperm cells (Agbaje et al. 2007; Navarro-Casado et al. 2010).

Considering the complex selective signatures of *WFDC8* and *SPINT4* (which might still represent recent adaptations), we propose that the selection acting on *WFDC8* and *SPINT4* may be related to innate immune functions in the reproductive tract, with possible consequences for fertility levels. This hypothesis is easier to reconcile with the geographic restriction of selective signatures that could be correlated with host-pathogen interaction and with the pathogen load, which largely differs in genus and number across world geographic regions (Prugnolle et al. 2005; Fumagalli et al. 2009).

We hypothesize that the *WFDC8* candidate variant −44A abolishes the binding sites of two transcription factors, possibly downregulating gene expression. This could provide a selective advantage related to augmentation of proteolytic activity in male secretions, which ultimately may

facilitate sperm cell capacitation and gain of motility. Alternatively, this variant might contribute to lower anti-inflammatory and antibacterial properties of male secretions, increasing the risk of developing urogenital infections. Given the much higher pathogen burden in Africa compared with Europe, the lower frequencies of −44A variant in the YRI sample and the geographic specificity of the selective signature to the CEU sample might be explained by differential allele fitness, subject to environmental and ecological variables.

The same adaptive pressure, driven by host-pathogen interactions, may drive selection of the *SPINT4* haplotype configuration [Ser73+98A] close to fixation in the YRI population. Here, joining the two alleles in a haplotype configuration—one leading to a modified Kunitz domain (of reduced inhibitory activity) and the other maintaining an active transactivator binding site—may increase the innate immunity function without compromising the proteolytic features (necessary for fertility) of male secretions. Exploring these hypotheses will clearly require functional characterization of the candidate variants; but the assumption of a fluid balance between the activities of *WFDCs* in regulating fertility and innate immunity is a plausible fit for the complex evolutionary scenarios evoked for *WFDC8* and *SPINT4*.

In conclusion, the links between *WFDC8* and *SPINT4* and complex selective scenarios probably reflect their concomitant key roles in reproduction and innate immunity. Functional characterization of the proposed selected variants −44A (*WFDC8*) and [Ser73+98A] (*SPINT4*) will be fundamental to gaining greater understanding of the evolutionary forces driving their evolution as well as their contribution to male fertility. Additional primate studies on the intraspecific diversity of these and other *WFDC* genes will contribute importantly to our understanding of their critical roles in reproduction and immunity.

Supplementary Material

Supplementary table S1, files 1 and 2, and figures 1–5 are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

We would like to thank Aida Andres, Max Planck Institute, for critically reading the manuscript and for insightful discussions; Jacquelyn K. Beals for the editing; and the two anonymous reviewers for their helpful feedback. This research was supported by the Portuguese Foundation for Science and Technology (FCT), project grant to S. Seixas—PTDC/SAU-GMG/64043/2006. Z. Ferreira is supported by SFRH/BD/45907/2008 fellowship from FCT, supported by POPH-QREN—Promotion of scientific employment, supported by the European Social Fund and national funds of the Ministry of Science, Technology and Higher Education. S. Seixas is supported by POPH-QREN—Promotion of scientific employment, supported by the European Social Fund and national funds of the Ministry of Science, Technology and Higher Education. IPATIMUP is an Associate Laboratory of the Portuguese Ministry of Science, Technology and Higher Education and is partially supported by FCT.

References

Agbaje IM, Rogers DA, McVicar CM, McClure N, Atkinson AB, Mallidis C, Lewis, SE. 2007. Insulin dependant diabetes mellitus: implications for male reproductive function. *Hum Reprod.* 22:1871–1877.

Akey JM. 2009. Constructing genomic maps of positive selection in humans: where do we go from here? *Genome Res.* 19:711–722.

Biswas S, Akey JM. 2006. Genomic insights into positive selection. *Trends Genet.* 22:437–446.

Biswas S, Scheinfeldt LB, Akey JM. 2009. Genome-wide insights into the patterns and determinants of fine-scale population structure in humans. *Am J Hum Genet.* 84:641–650.

Charlesworth D. 2006. Balancing selection and its effects on sequences in nearby genome regions. *PLoS Genet.* 2:e64.

Chevin LM, Hospital F. 2008. Selective sweep at a quantitative trait locus in the presence of background genetic variation. *Genetics* 180:1645–1660.

Chimpanzee Sequencing and Analysis Consortium. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Clauss A, Lilja H, Lundwall A. 2005. The evolution of a genetic locus encoding small serine proteinase inhibitors. *Biochem Biophys Res Commun.* 333:383–389.

Coop G, Griffiths RC. 2004. Ancestral inference on gene trees under selection. *Theor Popul Biol.* 66:219–232.

Crawford DC, Akey DT, Nickerson DA. 2005. The patterns of natural variation in human genes. *Annu Rev Genomics Hum Genet.* 6:287–312.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nat Genet.* 36:1326–1329.

Edstrom AM, Malm J, Frohm B, Martellini JA, Giwercman A, Morgelin M, Cole AM, Sorensen OE. 2008. The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins. *J Immunol.* 181:3413–3421.

Excoffier L. 2002. Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev.* 12:675–682.

Fay JC, Wu CI. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

Frisse L, Hudson RR, Bartoszewicz A, Wall JD, Donfack J, Di Rienzo A. 2001. Gene conversion and different population histories may explain the contrast between polymorphism and linkage disequilibrium levels. *Am J Hum Genet.* 69:831–843.

Fu YX, Li WH. 1993. Statistical tests of neutrality of mutations. *Genetics* 133:693–709.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* 19:199–212.

Garrigan D, Hammer MF. 2006. Reconstructing human origins in the genomic era. *Nat Rev Genet.* 7:669–680.

Gordo I, Gomes MG, Reis DG, Campos PR. 2009. Genetic diversity in the SIR model of pathogen evolution. *PLoS One.* 4:e4876.

Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344:403–410.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5:e1000695.

Hayakawa T, Aki I, Varki A, Satta Y, Takahata N. 2006. Fixation of the human-specific CMP-N-acetylneuraminic acid hydroxylase pseudogene and implications of haplotype diversity for human evolution. *Genetics* 172:1139–1146.

Hey J, Harris E. 1999. Population bottlenecks and patterns of human polymorphism. *Mol Biol Evol.* 16:1423–1426.

Hudson RR. 2001. Two-locus sampling distributions and their application. *Genetics* 159:1805–1817.

Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.

Hurle B, Swanson W, Program NCS, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. *Genome Res.* 17:276–286.

Kim HL, Igawa T, Kawashima A, Satta Y, Takahata N. 2010. Divergence, demography and gene loss along the human lineage. *Philos Trans R Soc Lond B Biol Sci.* 365:2451–2457.

Kim HL, Satta Y. 2008. Population genetic analysis of the N-acylsphingosine amidohydrolase gene associated with mental activity in humans. *Genetics* 178:1505–1515.

Kingan SB, Tatar M, Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *J Mol Evol.* 57:159–169.

Lapidot M, Mizrahi-Man O, Pilpel Y. 2008. Functional characterization of variations on regulatory motifs. *PLoS Genet.* 4:e1000018.

Lundwall A. 2007. A locus on chromosome 20 encompassing genes that are highly expressed in the epididymis. *Asian J Androl.* 9:540–544.

Lundwall A, Clauss A, Olsson AY. 2006. Evolution of kallikrein-related peptidases in mammals and identification of a genetic locus encoding potential regulatory inhibitors. *Biol Chem.* 387:243–249.

Macedo-Ribeiro S, Almeida C, Calisto BM, Friedrich T, Mentele R, Sturzebecher J, Fuentes-Prior P, Pereira PJ. 2008. Isolation, cloning and structural characterisation of boophilin, a multifunctional Kunitz-type proteinase inhibitor from the cattle tick. *PLoS One.* 3:e1624.

Marinescu VD, Kohane IS, Riva A. 2005a. MAPPER: a search engine for the computational identification of putative transcription factor binding sites in multiple genomes. *BMC Bioinformatics.* 6:79.

Marinescu VD, Kohane IS, Riva A. 2005b. The MAPPER database: a multi-genome catalog of putative transcription factor binding sites. *Nucleic Acids Res.* 33:D91–D97.

McCrudden MT, Dafforn TR, Houston DF, Turkington PT, Timson DJ. 2008. Functional domains of the human epididymal protease inhibitor, eppin. *FEBS J.* 275:1742–1750.

Murer V, Spetz JF, Hengst U, Altrogge LM, de Agostini A, Monard D. 2001. Male fertility defects in mice lacking the serine protease inhibitor protease nexin-1. *Proc Natl Acad Sci U S A.* 98:3029–3033.

Navarro-Casado L, Juncos-Tobarra MA, Chafer-Rudilla M, Iniguez-de Onzono L, Blazquez-Cabrera JA, Miralles-Garcia JM. 2010. Effect of experimental diabetes and STZ on male fertility capacity: study in rats. *J Androl.* 31:584–592.

Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25:2745–2751.

O'Rand MG, Widgren EE, Sivashanmugam P, Richardson RT, Hall SH, French FS, VandeVoort CA, Ramachandra SG, Ramesh V, Jagannadha Rao A. 2004. Reversible immunocontraception in male monkeys immunized with eppin. *Science* 306:1189–1190.

O'Rand MG, Widgren EE, Wang Z, Richardson RT. 2007. Eppin: an epididymal protease inhibitor and a target for male contraception. *Soc Reprod Fertil Suppl.* 63:445–453.

Patin E, Harmant C, Kidd KK, Kidd J, Froment A, Mehdi SQ, Sica L, Heyer E, Quintana-Murci L. 2006. Sub-Saharan African coding sequence variation and haplotype diversity at the NAT2 gene. *Hum Mutat.* 27:720.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.

Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 15:1022–1027.

Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59:2312–2323.

Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. *Nucleic Acids Res.* 30:3894–3900.

Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Mol Biol Evol.* 25:207–219.

Rozas J. 2009. DNA sequence polymorphism analysis using DnaSP. *Methods Mol Biol.* 537:337–350.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Sabeti PC, Walsh E, Schaffner SF, et al. (15 co-authors). 2005. The case for selection at CCR5-Delta32. *PLoS Biol.* 3:e378.

Sallenave JM. 2010. Secretory leukocyte protease inhibitor and elafin/trappin-2: versatile mucosal antimicrobials and regulators of immunity. *Am J Respir Cell Mol Biol.* 42:635–643.

Satta Y, Takahata N. 2004. The distribution of the ancestral haplotype in finite stepping-stone models with population expansion. *Mol Ecol.* 13:877–886.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Scheinfeldt LB, Biswas S, Madeoy J, Connelly CF, Schadt EE, Akey JM. 2009. Population genomic analysis of ALMS1 in humans reveals a surprisingly complex evolutionary history. *Mol Biol Evol.* 26:1357–1367.

Seixas S, Suriano G, Carvalho F, Seruca R, Rocha J, Di Rienzo A. 2007. Sequence diversity at the proximal 14q32.1 SERPIN subcluster: evidence for natural selection favoring the pseudogenization of SERPINA2. *Mol Biol Evol.* 24:587–598.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Teshima KM, Coop G, Przeworski M. 2006. How reliable are empirical genomic scans for selective sweeps? *Genome Res.* 16:702–712.

Thimon V, Calvo E, Koukoui O, Legare C, Sullivan R. 2008. Effects of vasectomy on gene expression profiling along the human epididymis. *Biol Reprod.* 79:262–273.

Tishkoff SA, Verrelli BC. 2003. Patterns of human genetic diversity: implications for human evolutionary history and disease. *Annu Rev Genomics Hum Genet.* 4:293–340.

Todd JA, Walker NM, Cooper JD, et al. 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet.* 39:857–864.

Uhrin P, Dewerchin M, Hilpert M, et al. 2000. Disruption of the protein C inhibitor gene results in impaired spermatogenesis and male infertility. *J Clin Invest.* 106:1531–1539.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A*. 102:18508–18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol*. 4:e72.

Wang ET, Kodama G, Baldi P, Moyzis RK. 2006. Global landscape of recent inferred Darwinian selection for Homo sapiens. *Proc Natl Acad Sci U S A*. 103:135–140.

Wang Z, Widgren EE, Richardson RT, O'Rand MG. 2007a. Characterization of an eppin protein complex from human semen and spermatozoa. *Biol Reprod*. 77:476–484.

Wang Z, Widgren EE, Richardson RT, O'Rand MG. 2007b. Eppin: a molecular strategy for male contraception. *Soc Reprod Fertil Suppl*. 65:535–542.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol*. 7:256–276.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174:1431–1439.

Zhu J, Nathan C, Jin W, et al. (11 co-authors). 2002. Conversion of proepithelin to epithelins: roles of SLPI and elastase in host defense and wound repair. *Cell* 111:867–878.

3.1 Evolutionary history of WFDCs in human populations

3.1.2 Reproduction and Immunity Driven Natural Selection in the Human *WFDC* Locus.

Mol Biol Evol. 2013 Apr;30(4):938-50

Reproduction and Immunity-Driven Natural Selection in the Human *WFDC* Locus

Zélia Ferreira,*[1,2,3] Susana Seixas,[2] Aida M. Andrés,[4] Warren W. Kretzschmar,[5] James C. Mullikin,[6] Praveen F. Cherukuri,[6] Pedro Cruz,[6] Willie J. Swanson,[7] NISC Comparative Sequencing Program,[1,6] Andrew G. Clark,[8] Eric D. Green,[1] and Belen Hurle[1]

[1]National Human Genome Research Institute, National Institutes of Health, Bethesda, Maryland

[2]Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto, Portugal

[3]Department of Biology, Faculty of Sciences, University of Porto, Porto, Portugal

[4]Department of Evolutionary Genetics, Max Planck Institute for Evolutionary Anthropology, Leipzig, Germany

[5]Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom

[6]NIH Intramural Sequencing Center (NISC), National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland

[7]Department of Genome Sciences, University of Washington

[8]Department of Molecular Biology and Genetics, Cornell University

***Corresponding author:** E-mail: zferreira@ipatimup.pt.

Associate editor: Ryan Hernandez

Abstract

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) locus located on human chromosome 20q13 spans 19 genes with WAP and/or Kunitz domains. These genes participate in antimicrobial, immune, and tissue homoeostasis activities. Neighboring *SEMG* genes encode seminal proteins Semenogelin 1 and 2 (SEMG1 and SEMG2). *WFDC* and *SEMG* genes have a strikingly high rate of amino acid replacement (d_N/d_S), indicative of responses to adaptive pressures during vertebrate evolution. To better understand the selection pressures acting on *WFDC* genes in human populations, we resequenced 18 genes and 54 noncoding segments in 71 European (CEU), African (YRI), and Asian (CHB + JPT) individuals. Overall, we identified 484 single-nucleotide polymorphisms (SNPs), including 65 coding variants (of which 49 are nonsynonymous differences). Using classic neutrality tests, we confirmed the signature of short-term balancing selection on *WFDC8* in Europeans and a signature of positive selection spanning genes *PI3*, *SEMG1*, *SEMG2*, and *SLPI*. Associated with the latter signal, we identified an unusually homogeneous-derived 100-kb haplotype with a frequency of 88% in Asian populations. A putative candidate variant targeted by selection is Thr56Ser in *SEMG1*, which may alter the proteolytic profile of SEMG1 and antimicrobial activities of semen. All the well-characterized genes residing in the *WDFC* locus encode proteins that appear to have a role in immunity and/or fertility, two processes that are often associated with adaptive evolution. This study provides further evidence that the *WFDC* and *SEMG* loci have been under strong adaptive pressure within the short timescale of modern humans.

***Key words:* WFDC, semenogelins, natural selection, innate immunity, serine protease inhibitors, reproduction.**

Introduction

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) gene locus on human chromosome 20q13 spans 19 genes with WAP and/or Kunitz domains that confer serine protease inhibitor activity (Clauss et al. 2005, 2011; Lundwall 2007; Lundwall and Clauss 2011). *WFDC* genes exhibit core functions involving reproduction, antimicrobial, immune, and tissue homoeostasis activities that in most cases remain poorly understood (Yenugu et al. 2004; Bouchard et al. 2006; Bingle and Vyakarnam 2008; Lundwall and Clauss 2011). The *WFDC* locus includes genes encoding the seminal proteins Semenogelin 1 and 2 (SEMG1 and -2) (Peter et al. 1998; de Lamirande 2007; Lundwall 2007). The *WFDC* and *SEMG* genes stand out for reports of striking signatures of adaptive evolution, reflecting effects of natural selection during mammalian evolution (Dorus et al. 2004; Hurle et al. 2007).

Most evolutionary and functional studies on the *WFDC* gene family have focused on genes located within the centromeric sublocus of the large gene cluster (fig. 1*A*). This small but dynamic genome region (hereafter referred to as *WFDC*-CEN) has a notably complex evolutionary history resulting in rapid interspecies divergence of both coding and noncoding sequences (Hurle et al. 2007). Proteins encoded by the genes in *WFDC*-CEN include the well-studied peptidase inhibitor 3 (PI3, also known as elafin) and secretory leucocyte proteinase inhibitor (SLPI), which are pleiotropic molecules engaged in the surveillance against microbial infections at mucosal surfaces (Williams et al. 2006; Weldon and Taggart 2007; Weldon et al. 2007; McKiernan et al. 2011). Also well characterized are the *SEMG1* and *SEMG2* genes encoding seminal plasma proteins with roles in semen clotting and in antimicrobial protection for the spermatozoa in the female reproductive tract

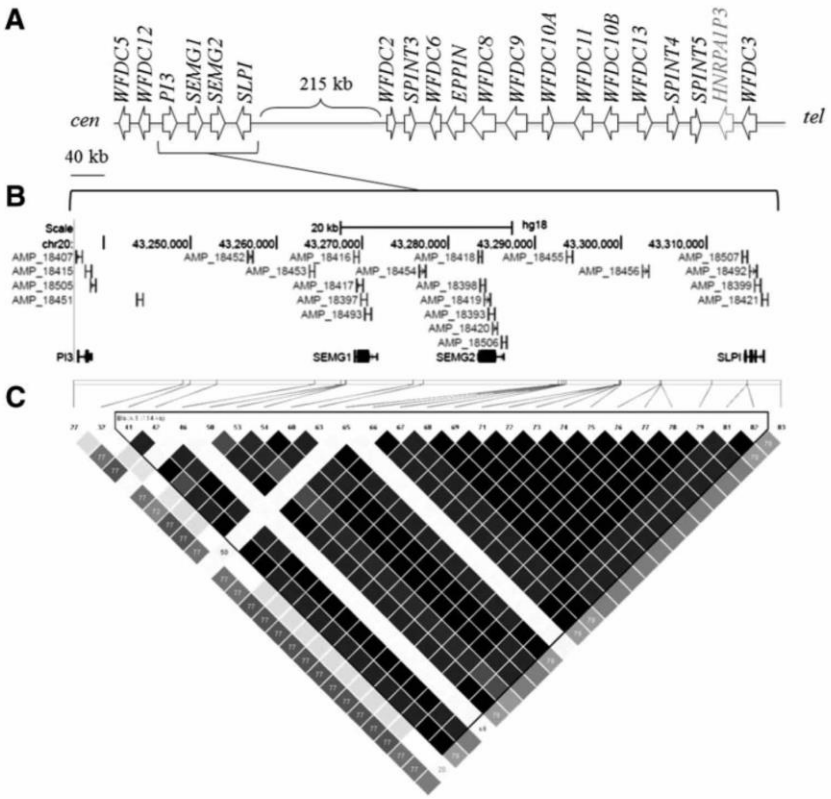


FIG. 1. Schematic representation of the 20q13 *WFDC* gene cluster. (*A*) Diagram showing the relative positions of the *WFDC* genes. As depicted, the *WFDC* cluster spans 700 kb and its genes are organized into two subloci (centromeric and telomeric; *WFDC*-CEN and *WFDC*-TEL, respectively), separated by 215 kb of unrelated sequence. *HNRPA1P3* pseudogene is indicated in light gray. (*B*) Strategy for the resequencing effort across the *WFDC* locus. One hundred thirty 700-bp-long amplicons were designed to include all exonic regions and a selection of noncoding sequences evenly spaced every 10 kb across the two *WFDC* subloci. (*C*) Linkage disequilibrium in the *PI3-SEMG1-SEMG2-SLPI* region in Asians (calculated using the resequenced data and displayed with Haploview; Haplotype blocks defined by Gabriel et al. 2002).

(Lundwall et al. 2002; Bourgeon et al. 2004; Edstrom et al. 2008; Martellini et al. 2009).

Comparative genomics and phylogenetic analysis indicate that *SLPI*, *PI3*, *SEMG1*, and *SEMG2* have evolved rapidly since the separation of the primate and murine lineages (Hurle et al. 2007). In particular, multiple studies show their accelerated molecular evolution as measured by their high d_N/d_S values (Dorus et al. 2004; Hurle et al. 2007; Ramm et al. 2008).

The *WFDC* telomeric sublocus (hereafter referred to as *WFDC*-TEL) is physically separated from *WFDC*-CEN by 215 kb of unrelated genomic sequence. The best-characterized gene within *WFDC*-TEL encodes the epididymal protease inhibitor (EPPIN, also known as *SPINLW1*). At ejaculation, EPPIN coats the surface of human spermatozoa, binds to SEMG1, and helps to modulate the activity of prostate-specific antigen (PSA) while providing antimicrobial protection for spermatozoa (Robert and Gagnon 1999; Bourgeon et al. 2004; Wang et al. 2005; Edstrom et al. 2008; Zhao et al. 2008). The functions of most other *WFDC* genes remain poorly characterized.

Surprisingly, despite the strong signatures of positive selection revealed by excess nonsynonymous (NS) divergence among species, few studies have used intraspecific polymorphism data to examine the selective pressures acting on *WFDC* and *SEMG* genes within populations. Most of these focused on *WFDC*-CEN, where the *SEMGs* have been identified as genes under adaptive evolution, specifically, by correlating their single-nucleotide polymorphisms (SNPs) and copy-number variants to the different mating systems of various primate species (Jensen-Seaman and Li 2003; Kingan et al. 2003; Dorus et al. 2004; Carnahan and Jensen-Seaman 2008).

The only study examining the selective pressures occurring in *WFDC*-TEL focused on *WFDC8*, a proposed target of recent balancing selection in Europeans based on the intermediate frequency of the haplotypes containing the candidate variant [-44(G/A)] and *SPINT4*, which was linked to a rapid increase in frequency of the advantageous allele (Ser73), associated with a long-range haplotype (LRH) and low-frequency variants—thus appearing to evolve under an incomplete selective sweep in Africans (Ferreira et al. 2011).

To gain a better understanding of the more recent selective pressures shaping genetic variation in the human *WFDC* locus, we systematically resequenced 18 genes of the locus plus 54 evenly spaced noncoding segments in 71 humans

from European (CEU), African (YRI), and Asian (CHB + JPT) HapMap populations. A set of 47 autosomal, unlinked, and neutrally evolving loci were also surveyed to assess baseline (neutral) genomic diversity. Using classic neutrality tests (Tajima's *D* and Fay and Wu's *H*), we confirmed the signature of short-term balancing selection on *WFDC8* in the CEU population; and we further pinpointed a signature of positive selection spanning *PI3*, *SEMG1*, *SEMG2*, and *SLPI*. The best candidate variant for the latter selective footprint in Asians was allele Ser56 in SEMG1. This variant potentially modifies the likelihood of PSA-mediated hydrolysis of SEMG1, simultaneously altering the peptide profile and antimicrobial activities of semen.

This study is the first to provide systematic and comprehensive population genomics-based evidence that a number of *WFDC* and *SEMG* genes are under strong adaptive pressures within the recent timescale of modern humans.

Results

To gain a better understanding of the selective pressures shaping the genetic variation within *WFDC* genes, we designed 130 (~700 bp) amplicons across the *WFDC* locus. These amplicons were amplified from a panel of 71 HapMap Phase I/II individuals (21 CEU, 25 YRI, and 25 CHB + JPT) and Sanger sequenced (supplementary tables S1 and S2, Supplementary Material online). In this study, a total of 8.1 Mb of targeted genomic regions were sequenced, 20% of which corresponds to exonic regions and the rest accounts for intronic and putative *cis*-regulatory regions (52%) and intergenic regions (28%) (supplementary table S3, Supplementary Material online).

Genetic Variation in WFDC Genes

Overall, 484 SNPs were identified, of which 65 resided in coding regions. Forty-nine of the coding SNPs were NS, of which 67% were present at very low frequencies in all populations ($f \leq 0.08$) (fig. 2; supplementary table S3a, Supplementary Material online). Such a pattern of allele frequencies is consistent with mildly deleterious effects of most NS variants, although it does not depart from a strictly neutral site frequency spectrum (SFS; 1,000 coalescent simulations; $S = 49$; χ^2 test; $P = 0.47$). Seven NS-SNPs were predicted to affect protein function by SIFT and PolyPhen v2 where only rs6017667 (Gly73Ser in SPINT4) occurs at an intermediate frequency $f = 0.44$.

Twenty-four insertions/deletions (indels) were found, 21 of which were located in intronic and intergenic regions. The three remaining indels were in untranslated coding regions of *WFDC9* and *WFDC13*. Because indels might have a distinct mutation rate compared with SNPs and their genomic localization does not seem to affect protein function or expression, they were excluded from the following analyses. Additionally, we found 456 fixed human–chimpanzee differences, of which only 19 were within coding regions and human specific. The Polyphen v2 and SIFT analysis show that the functional impact of most of the NS fixed differences was classified as benign (supplementary table S3b, Supplementary Material online).

Deviations of Allele Frequency Spectra from Neutral Expectations

Figure 2 depicts the distribution of folded SFS for all the surveyed genes. In *WFDC*-CEN, there is a skew toward low-frequency variants (fig. 2*A* and *B*), whereas in *WFDC*-TEL, there is a shift toward intermediate-frequency variants (fig. 2*C* and *D*), following the trend observed for coding SNPs. The significance of deviations from neutral expectations of the SFS within each population was tested using the summary statistics π, θ_w, Tajima's *D*, and Fay and Wu's *H* (Tajima 1989; Fu 1996; Fay and Wu 2000; Zeng et al. 2006). We controlled for demography effects by using the demographic model developed by Gutenkunst et al. (2009) and determined nominal *P* values for each statistic (table 1 and supplementary table S4, Supplementary Material online). Individual genes in the *WFDC* locus present summary statistics that have moderate *P* values and whose significance is marginal after multiple test correction (Benjamini and Hochberg 1995; Storey 2002; Storey and Tibshirani 2003; Storey et al. 2004). However, the nominal *P* values clearly show that there is a trend toward lower nominal *P* values for Tajima's *D*, Fu and Li's *D*, Fay and Wu's *H*, and Mann-Whitney U $(MWU)_{high}$ (supplementary fig. S2, Supplementary Material online), which suggested the need for further testing.

The tail probability of test statistics of the *WDFC* region was assessed by using simulations based on fits of demographic models to the neutral regions. One evaluation of the validity of this approach is to determine the corresponding tail probabilities for the control regions in the study. We calculated the levels of nucleotide diversity (π) and Tajima's *D* for the 47 control regions in each population and created an empirical distribution of the obtained values. Because the control regions have been subject to the same demographic history as the *WFDC* locus, an outlier value (2.5 or 97.5 percentile) would suggest a non-neutral evolution event (supplementary fig. S3, Supplementary Material online).

At the population level, the lowest π levels were found mainly in the Asian population, followed by the CEU and YRI populations (supplementary table S4, Supplementary Material online), as expected under the out-of-Africa model for human populations (Schaffner et al. 2005; Voight et al. 2005; Gutenkunst et al. 2009). At the gene level, the genes that display the most unusual π values (supplementary fig. S3*A* and table S4, Supplementary Material online) are *SEMG1* and *SEMG2*, with low nucleotide diversity values ($\pi_{SEMG1} = 0.761063 \times 10^{-4}$; $\pi_{SEMG2} = 0.933816 \times 10^{-4}$) in the Asian population, and *WFDC3*, with high nucleotide diversity in Europeans and Africans ($\pi_{WFDC3} = 11.473 \times 10^{-4}$ and $\pi_{WFDC3} = 14.0656 \times 10^{-4}$ for each population, respectively). The generated empirical distribution of Tajima's *D* values compared with each gene suggests that *PI3* and *SEMG2* are outliers in the Asian population (supplementary fig. S3*B*, Supplementary Material online). The overall levels of diversity

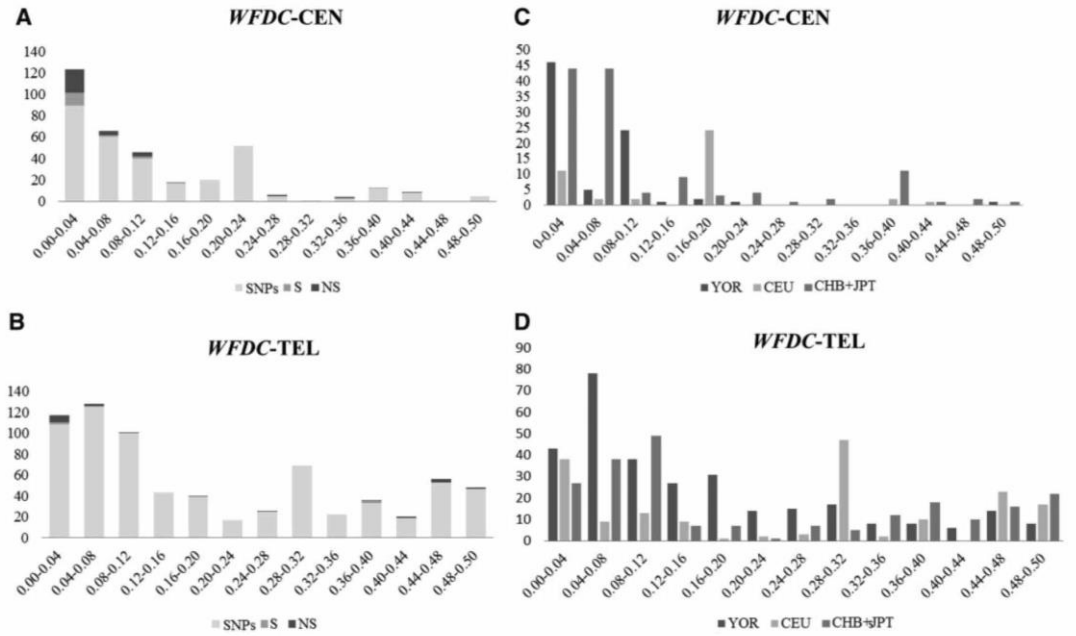


FIG. 2. Folded site frequency spectrum (folded SFS) for the *WFDC* in all populations resequenced. The *x* axis depicts the frequency of the allele frequency bin in the generated data set, whereas the *y* axis represents the number of alleles found within each frequency bin. S, synonymous changes; NS, nonsynonymous changes. (*A*) and (*B*), folded SFS in *WFDC*-CEN; (*C*) and (*D*), folded SFS in *WFDC*-TEL.

Table 1. Significant Summary Statistics at the *WFDC* Locus.

Gene	Population	S[a]	θ_w[b]	π[c]	Tajima's *D*[d]	*H*[e]
PI3	YRI	34	7.60101	4.69663	−1.28656	−1.70946**
	CHB + JPT	17	3.81646	1.67945	−1.75823*	−3.79418**
SEMG2	YRI	18	4.02479	2.81504	−0.947299	−1.68143**
	CHB + JPT	11	2.46906	0.933816	−1.82029*	−3.54599**
SLPI	CEU	10	2.56518	1.86032	−0.884379	−2.60125*
	CHB + JPT	16	3.59903	2.76312	−0.724534	−2.85197*
SPINT3	YRI	11	2.45787	1.32809	−1.33954	−1.60077**
WFDC8	CEU	27	6.88872	10.7402	2.02506*	−0.250797
	CHB + JPT	31	7.01214	5.01087	−0.964878	−3.7064**

[a]S, number of segregating sites.
[b]Watterson's estimator of θ ($4N_e\mu$) (Watterson 1975) per base pair ($\times 10^{-4}$).
[c]Nucleotide diversity per base pair ($\times 10^{-4}$).
[d]Tajima's *D* statistic (Tajima 1989).
[e]Fay and Wu's *H* test (Fay and Wu 2000; Zeng et al. 2006).
*$P < 0.05$ and **$P < 0.025$.

in the *WFDC* locus suggest a non-neutral evolution of these genes.

Footprints of Recent Positive Selection in Asians

Summary statistics suggest that the *PI3-SEMG1-SEMG2-SLPI* region at *WFDC*-CEN has evolved under non-neutral evolution in the Asian population (table 1 and supplementary table S4, Supplementary Material online). Considering the physical clustering of these genes (fig. 1), their low levels of intrapopulation nucleotide diversity, and outlier Tajima's *D* values both in the empirical and simulated comparisons, we looked for possible signatures of positive selection in this region.

Considering each gene individually, these have borderline significant Tajima's *D* and Fay and Wu's *H P* values (table 1). Additionally, a number of SNPs in *SEMG1* and *SLPI* presented elevated F_{ST} values, with *P* values ranging from 0.01 to 0.05 (supplementary fig. S1*A*–*F*, Supplementary Material online). Coincidentally, the *PI3-SEMG1-SEMG2-SLPI* (fig. 1*A*) gene array forms a single linkage disequilibrium (LD) block ($D' = 1$; $r^2 = 0.8$) of approximately 100 kb (fig. 1*B* and *C*) in the Asian population. This LD block is longer than the average Asian LD extension of approximately 44 kb and longer than the flanking regions (Gabriel et al. 2002).

To determine whether the low levels of diversity and elevated haplotype extension could be due to natural selection,

we used a sliding window of Tajima's *D* with the aim to identify peaks of aberrant Tajima's *D* values across this region (fig. 3). Two peaks of very low Tajima's *D* were identified: one between *PI3* and *SEMG1* (−2.17) and another one in *SEMG2* (−1.86). Downstream of *SEMG2*, the levels of diversity variation recover rapidly, and *SLPI* no longer has low levels of diversity, with higher Tajima's *D*, π, and haplotype diversity values than its neighboring genes. These results suggest that *SLPI* is not a gene under selective pressure.

Taking into account the strong LD among these genes, we sought for signals of recent selection using the following haplotype tests: Hudson's haplotype test (Hudson et al. 1994), derived intra-allelic nucleotide diversity (DIND; Barreiro et al. 2009), and extended haplotype homozygosity (EHH) and relative EHH (REHH) (Sabeti et al. 2002). We started by applying the Hudson's haplotype test because the LD block shown in figure 1C could be attributed to a single haplotype with a 100-kb extension present in 88% of the in Asian population. With such test, we examined whether the common haplotype contained fewer segregating sites than expected under neutrality given its frequency. We obtained significant results for the 100 kb region, encompassing 8 variable positions out of 44 segregating sites. The tests were based on 10,000 simulations (ms, Hudson 2002) of the constant neutral model ($P = 0.0023$) and on the Gutenkunst model for Asian populations (without migration) ($P = 0.0205$) (Gutenkunst et al. 2009).

In spite of the low nucleotide diversity associated with the extended haplotype, the DIND test showed that none of the *PI3*, *SEMG1*, or *SEMG2* variants was significant under the Gutenkunst demographic model. Nonetheless, a number of variants (rs13042431, rs2267864, rs2301366, and rs2071651; fig. 4) located in the region of interest display borderline nonsignificance under the Gutenkunst model and present significant values under the constant model of demography. From the latter SNPs, only one is a NS variant (rs2301366 A→T), reflecting an amino acid change of Thr56Ser in SEMG1 (fig. 4).

To further evaluate whether this haplotype structure could result from the action of positive selection, we calculated the EHH statistic proposed by Sabeti et al. (2002). We started by centering our analysis in the only NS variant that presented borderline *P* values from the DIND test, as a possible candidate variant of selection. Specifically, we measured the decay of LD around a three-SNP core haplotype centered in Thr56Ser. The bifurcation plot associated with Ser56 shows a frequent haplotype that extends for more than 60 kb in both directions from Thr56Ser (fig. 5A). We determined whether the EHH for the Ser56 core haplotype in the Asian sample was unusual by comparing its frequency and REHH at the largest distance where non−T/A haplotypes had nonzero values of EHH (80 kb distal and proximal) against null distributions. The deviations from simulated null distributions were significant for the haplotype associated with Ser56 in Asians (fig. 5B) but not in the other populations (results not shown).

A better understanding of the evolutionary history of *SEMG1* was obtained by the analysis of haplotype genealogies.

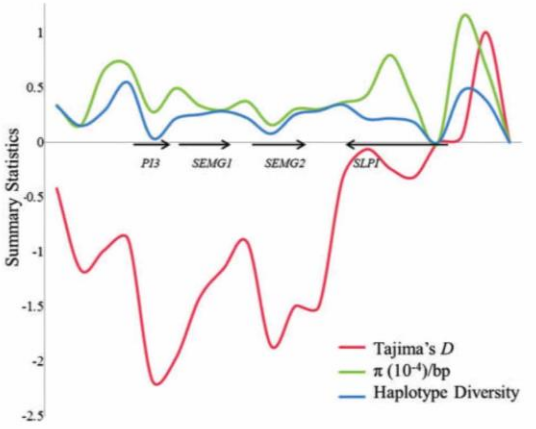


FIG. 3. Sliding window of Tajima's *D*, π, and Haplotype diversity (red, green, and blue lines, respectively) in the *PI3-SEMG1-SEMG2-SLPI* region in Asians. *PI3* and *SEMG1* region shows lower values than the rest of *WFDC*-CEN. Window size,1,000 bp; increment, 500 bp.

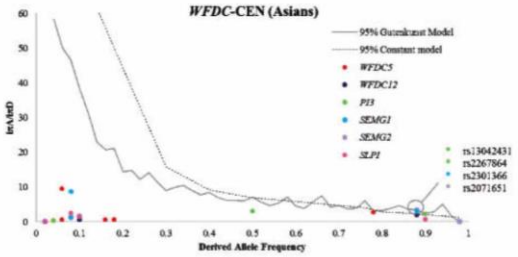


FIG. 4. Ratio of the ancestral ($i\tau_A$) alleles to the haplotypes carrying the derived ($i\tau_D$) alleles above expected, plotted as a function of Derived allele frequency. $P < 0.05$; Dashed line—5% constant model, recombination at 0.2 *c*M. Solid line—5% Gutenkunst model (Gutenkunst et al. 2009) in *WFDC*-CEN for Asians.

To estimate the divergence time of *SEMG1* and the age of Thr56Ser, we used a maximum likelihood coalescent analysis by GENETREE (Griffiths and Tavare 1994) and estimated the time to most recent common ancestor (T_{MRCA}). Because we suspected a selective force acting on Ser56, we estimated the β parameter and used it to determine gene trees under selection (Coop and Griffiths 2004). Using all three populations, the estimated T_{MRCA} for the entire *SEMG1* genealogy was 0.675 ± 0.103 My and for the Ser56 variant was 0.287 ± 0.05 My ($\theta_{ML} = 6.13$; $\beta = 1.70$). We then reconstructed the haplotype phylogenetic network of *SEMG1* using a median-joining algorithm (fig. 6A). Specifically, most Asian haplotypes cluster around a haplotype defined by Thr56Ser (rs2301366). The derived allele (T) is shared among all the descendent haplotypes, showing a star-shaped haplotype network, which is usually associated with a selective sweep or population expansion.

The haplotype tests and network phylogenetic structure suggest a non-neutral evolution of *SEMG1* (combined

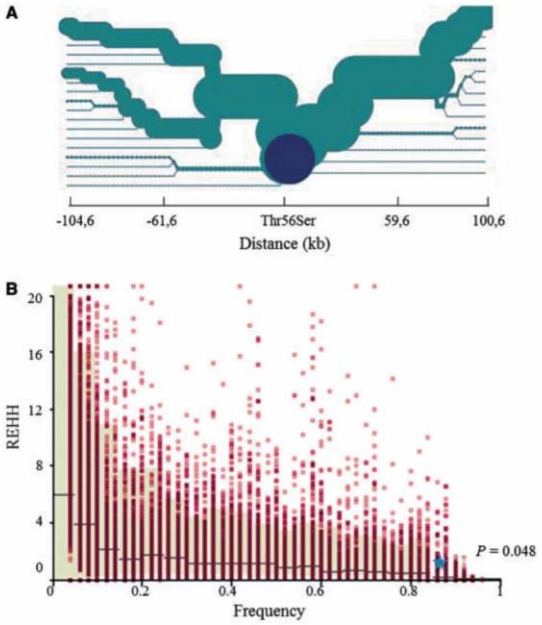


FIG. 5. (*A*) Haplotype bifurcation plot centered in position Thr56Ser of SEMG1 in Asian populations, using SWEEP. Thr56Ser is marked with a dark circle. The diameter of the circle and arm length is proportional to the number of individuals with the same LRH. Each of the additional SNPs is represented by a node from which bifurcation indicates a recombination event. (*B*) Relative expected haplotype homozygosity (REHH) deviations from simulated null distributions in the Asian population, using SWEEP software (www.broadinstitute.org/mpg/sweep, last accessed January 14, 2013). Highlighted point (star, P = 0.048) is Thr56Ser.

z-weighted *P* values = 0.002) (Whitlock 2005). We hypothesize that Ser56 is likely be under the influence of a selective sweep representing an advantageous allele that was swept to higher frequency in the Asian population (88%), simultaneously lowering the overall levels of nucleotide diversity and increasing the haplotype homozygosity in 160 kb of the surrounding regions (*PI3-SEMG1-SEMG2-SLPI*). However, because of the less recent age of the candidate variant and its presence in the other sequenced populations at somewhat elevated frequencies (Europeans 80% and Africans 38%), one cannot rule out the possibility of an event of selection on standing variation of Ser56 (Przeworski et al. 2005; Pritchard et al. 2010; Hernandez et al. 2011).

We set to compare our summary statistics (based on Sanger validated data from the *WFDC* locus) with summary statistics generated from data obtained from human genetic variation in public reference projects. The goal was to evaluate whether the different sequencing methods used had an effect in detecting genomic outliers, potentially affected by non-neutral evolution. Specifically, we performed a principal component and SFS analyses for the 1000 Genomes Project (supplementary figs. S4 and S6*B* and *D*, Supplementary Material online; Patterson, Price, et al. 2006) and generated SFS for the Complete Genomics Diversity Panel (Complete Genomics Assembly v1.3; Drmanac et al. 2010). However, given the substantial differences in SNP distribution in the latter data set due to the low sample size per population (supplementary fig. S6*A* and *C*, Supplementary Material online), we chose in the analysis that follows to directly compare the variants found in the 1000 Genomes Project and our sequencing survey, restricting our attention only to the sample that was sequenced in both projects.

For the *WFDC* Locus, our Sanger-based sequencing strategy detected 80% of the SNPs gathered by the 1000 Genomes Project. Conversely, the 1000 Genomes data contains 75% of the SNPs present in the data set generated for this study. The bulk of the discrepancies lie in low-frequency variants for which the 1000 Genomes data set presents lower singleton, doubleton, and tripleton frequencies (supplementary fig. S7, Supplementary Material online). These findings show that the publicly available genomes are very useful to detect genomic outliers, even though they do not yet completely replace deeper coverage and high-quality sequencing data. Despite the differences between SFS for the *WFDC* locus, the summary statistics present very similar and relatable values (supplementary tables S5 and S6, Supplementary Material online) suggesting that both approaches lead to the same results in this region of the genome. Specifically, for *SEMG1* in the Asian population, the summary statistic values ($\pi_{SEMG1} = 0.805 \times 10^{-4}$; Tajima's $D = -2.07$; and Fu and Li's $D = -3.9752$) are consistent with a non-neutral evolution of this gene.

Footprint of Short-Term Balancing Selection in Europeans

A previous study indicated that *WFDC8* is under short-term balancing selection in the CEU population (Ferreira et al. 2011). Sequencing the entire *WFDC* locus in three HapMap populations provided an opportunity to test in a larger data set the selective signal centered on *WFDC8*. The resulting sequence data confirmed that *WFDC8* has a positive Tajima's *D* (2.02) and elevated π values (10.7×10^{-4}) in the CEU population (table 1). The folded SFS for *WFDC8* shows an excess of polymorphic sites with intermediate frequency (fig. 2*C* and *D* and supplementary fig. S6, Supplementary Material online), which is significant in the CEU population based on MWU_{high} test ($P = 0.0089$) (Nielsen et al. 2009; Andrés et al. 2010) (supplementary table S4, Supplementary Material online). The haplotype network of *WFDC8* is structured into two highly differentiated haplotypes: "Haplotype A" and "Haplotype B" both several mutations away from the ancestral state (fig. 6*B*). Furthermore, the analysis of the 1000 Genomes data set confirms the elevated Tajima's *D* value (2.11) of *WFDC8* in the CEU population (supplementary table S5, Supplementary Material online).

In combination, these results confirm that Haplotypes A and B differ at SNP rs7273669 (A/G), which is located 44 bp upstream the translation start site (hereafter, we refer to this SNP as -44(A/G) for simplicity) and presents elevated F_{ST} values in the European/Asian comparison ($F_{ST} = 0.52$; $P = 0.0026$; supplementary fig. S1*B*, Supplementary Material online). This SNP, situated in the 5′-region of *WFDC8*,

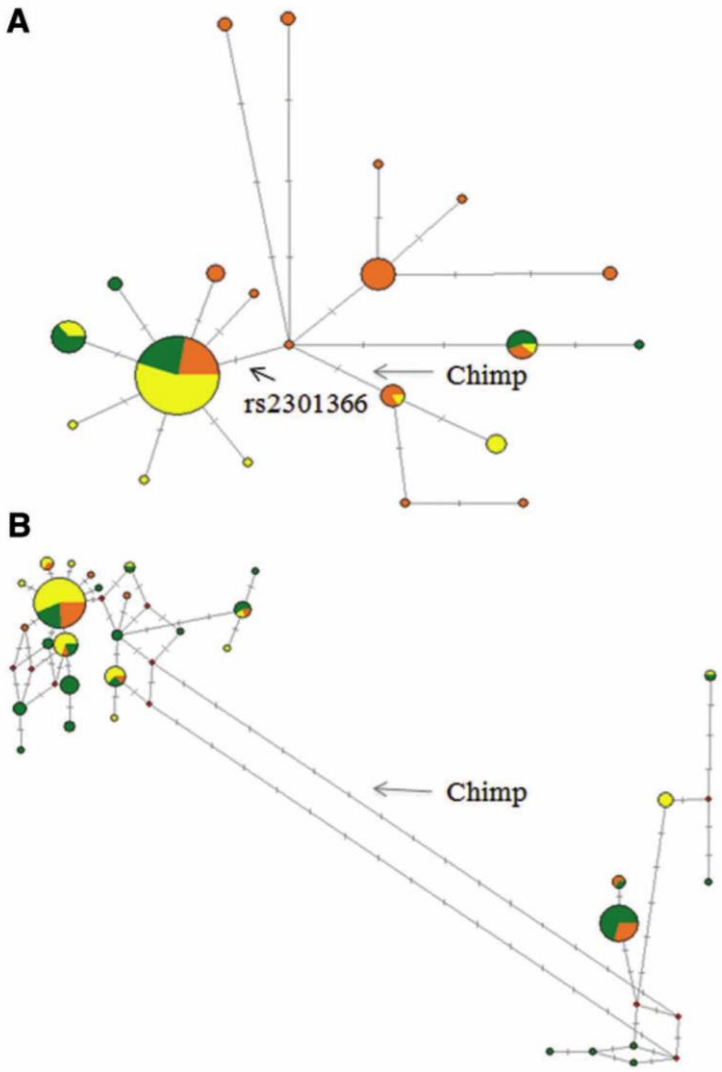


Fig. 6. Examples of inferred network haplotypes at the *WFDC* locus. Each circle represents a unique haplotype, and its area is proportional to its frequency. Within each circle, YRI, CEU, and Asian populations are labeled in orange, green, and yellow, respectively. The mutations that differentiate each haplotype are shown along each branch. The inferred network haplotype of *SEMG1* (*A*) and *WFDC8* (*B*) show a star-like structure (characteristic of a population expansion or recent selective sweep) and two highly differentiated haplotypes (characteristic of population structure or balancing selection), respectively.

potentially affects *cis*-regulatory elements that regulate *WFDC8* expression and has been proposed to affect the binding of two transcription factors (Ferreira et al. 2011).

Footprint of Incomplete Selective Sweep in Africans

We found other regions that stood out as possibly being under selection. A natural target of focused interest was *SPINT4* in *WFDC*-TEL, as this has been previously described as a candidate gene under selection in YRI populations based on the integrated Haplotype Score (Voight et al. 2006) and limited sequencing studies (Ferreira et al. 2011). *SPINT4* and the neighbor gene *WFDC3* do not show significant departure from the patterns expected under neutral evolution in the summary statistics, either in our sequencing study or the 1000 Genomes project data set (supplementary tables S4 and S5, Supplementary Material online), but they display elevated levels of F_{ST} (supplementary fig. S1*B* and *D*, Supplementary Material online) in African/non-African comparisons. Furthermore, variant rs6017667, a NS change that codes for Gly73Ser in *SPINT4*, stands out as the only NS-SNP with intermediate frequency (0.44) (supplementary table S3, Supplementary Material online) and elevated F_{ST} (F_{ST} (Af/As) = 0.26 [$P = 0.04$]; F_{ST} (Af/Eu) = 0.46 [$P = 0.04$]) (supplementary fig. S1*D* and *F*, Supplementary Material online), a signature that is typical of a variant under non-neutral evolution. *SPINT4* was previously thoroughly studied in YRI,

addressing significant EHH/REHH levels and a "star"-shaped genealogy typical of an ongoing positive selection event (Ferreira et al. 2011). Moreover, the Gly73Ser change, previously identified as the candidate variant of the incomplete selective sweep (Ferreira et al. 2011), is encoded by sequences within the second exon of *SPINT4*, which codes for the Kunitz domain and is, therefore, responsible for its serine protease inhibitor activity. Gly73Ser has been identified as a modification that affects stability of SPINT4 and has potential functional repercussions (supplementary table S3, Supplementary Material online).

Discussion

Previous efforts to identify targets of natural selection in the human genome have found an excess of selection acting on genes that mediate response to microbial attack and that play a role in reproduction. By studying the detailed patterns by which natural selection generates deviations from neutral evolution within the *WFDC* region, we can gain insights into the biological roles of specific *WFDC* and *SEMG* genes in host defense and reproduction. This effort to pinpoint signals of selective pressures shows a remarkable degree of interpopulation heterogeneity, identifying different genes under positive selection in three human populations. We hypothesize that this interpopulation heterogeneity is driven by the lack of homogeneity of pathogenic agents across the globe—however, further work will be required to identify agents specifically accounting for the specific patterns seen with the *WFDC* and *SEMG* genes.

The discovery of selective signals highlighted in this study takes into account that demographic history and genetic drift can affect both population differentiation. Although demographic processes affect the whole genome, natural selection acts on specific loci. Hence, the effect of demography must be controlled by comparing the genes of interest with an empirical distribution built from neutrally evolving regions of the genome. In this study, such control regions were represented by 47 unlinked, neutrally evolving pseudogenes (Andrés et al. 2010). We used a strict set of criteria to control the number of false positives and to maximize the detection of specific footprints of natural selection. In addition, because we performed various summary statistic-based tests to describe genetic variation across the *WFDC* locus, we corrected the *P* values for multiple testing by calculating *q* values from the obtained *P* values, estimating the proportion of false positives among the tests found to be significant (Benjamini and Hochberg 1995; Storey 2002; Storey and Tibshirani 2003; Storey et al. 2004). Notwithstanding, most studies presenting a candidate region approach usually present nominal *P* values when referring to a comparison with a particular demographic scenario (Barreiro et al. 2008, 2009; Fornarino et al. 2011; Hancock et al. 2011). Although the tests based on summary statistics failed to survive multiple test correction, they prompted us to pursue an analysis that combined the results of the different tests to probe distinct aspects of the data, including SNP allele frequencies, EHH, and population differentiation (F_{ST}).

Summary statistics, represented by Tajima's *D* and Fay and Wu's *H*′, suggest that *PI3*, *SEMG2*, and *SLPI* show a skew toward low-frequency variants in the Asian population, signals of a population expansion or positive selection. To discriminate between these possibilities, we performed coalescent simulations under a neutral demographic model (Gutenkunst et al. 2009) and compared Tajima's *D* statistic calculated for the sequenced neutrally evolving regions. The test results led us to conclude that *PI3*, *SEMG2*, and *SLPI* are not evolving under neutrality. In addition, the sliding window of Tajima's *D* performed in this region shows extremely negative values in *PI3* and *SEMGs*. Consistently, the summary statistics of these genes in the 1000 Genomes data set present low nucleotide diversity and strongly negative Tajima's *D* values, especially *SEMG1*, which presents the lowest values of the entire *WFDC* locus. These results point toward positive selection acting in this region.

Subsequent analysis of population differentiation of these loci found that some of the SNPs have elevated values of F_{ST}, suggesting the possibility that they might be under region-specific selective pressures. The single NS SNP among those with high F_{ST} values is rs2301366, a variant located on the second exon of *SEMG1* and responsible for the Thr56Ser replacement (ACC→TCC). The derived state of this variant is present in 88% of the Asian samples and defines a haplotype that spans 160 kb. The haplotype-based tests (Hudson haplotype test, DIND, and EHH/REHH) indicate that Ser56 haplotype has unusually low levels of intrahaplotypic diversity and long-range extension given its frequency, which significantly deviates from neutrality under the calibrated model of Asian demography.

Network analysis of the composite haplotypes for all populations suggests the Thr56Ser as the most plausible target of selection. Although the haplotype cladogram shows a star-like structure that can be characteristic of a population expansion, the previous tests performed suggest positive selection in the *PI3-SEMG1-SEMG2* region in Asians, centered on Thr56Ser. Ancestral Thr56 in SEMG1 is highly conserved among primates, dating back to Old World Monkeys, and conserved at position 56 of paralogous gene SEMG2, 79% similar in sequence to SEMG1 (Hurle et al. 2007). The derived allele, Ser56, is also present in African (38% frequency) and European (80% frequency) populations and consistently has a 0.287 Ma age estimate before the "Out-of-Africa" migrations. For the expectations of a classical selective sweep, Ser56 may have weak footprints as indicated also by the shorter haplotype and borderline summary statistics. Conversely, Ser56 may provide a good fit for a model of selection on standing variation in which an allele already segregating in a population is favored by a sudden change in selective pressures (Przeworski et al. 2005; Pritchard et al. 2010; Hernandez et al. 2011). Although the variant Ser56 does not seem to affect SEMG1 protein structure or stability, SEMG1 has a well-established role in forming semen coagulum, crosslinking with SEMG2 to entrap spermatozoa, priming them for optimal fertilization potential (sperm capacitation). Later, this coagulum is degraded by the action of PSA, which cleaves the crosslinking matrix, and releases spermatozoa along with SEMG1- and SEMG2-derived peptides. The N-terminal peptides from SEMG1, with a Serine in position 56, have been described

to have antimicrobial and antiviral activity both in male and female reproductive tracts, whereas the peptides originated from SEMG2, with a Threonine in position 56, do not present antimicrobial activity (Robert and Gagnon 1999; Bourgeon et al. 2004; Edstrom et al. 2008; Zhao et al. 2008; Martellini et al. 2009). The location of Ser56, six amino acids upstream of a mapped PSA cleavage site (Tyr63), may alter the efficacy of the cleavage at this site compared with other primates (Robert et al. 1997; Bourgeon et al. 2004). Here, we hypothesize that the change from Thr to Ser may change the proteolytic cleavage of SEMG1 by PSA, leading to a modified peptide profile and antimicrobial activities.

A signal of short-term balancing selection in Europeans centered in *WFDC8* and the incomplete selective sweep in *SPINT4* were previously described at the *WFDC* region in CEU and YRI, respectively (Ferreira et al. 2011). When the selective signal in *WFDC8* was re-examined in an independent sample, the remarkable differentiation between the Haplotypes A and B in CEU, along with the intermediate frequency at which they are found, solidifies the evidence that *WFDC8* is under a balancing selection in this population. Variant -44(A/G) remains the best candidate SNP under selection, potentially regulating the expression of *WFDC8*. Specifically, the intermediate frequency of these alleles may regulate the levels of *WFDC8* expression, maximizing its role in proteolysis cascades linked to sperm maturation, as well as its antimicrobial functions. In fact, Wfdc8 has been shown to have antibacterial activity in rat male reproductive tract (Rajesh et al. 2011), and its ortholog in humans, *WFDC8*, has been shown to be associated to impaired fertility (Thimon et al. 2008).

Similarly, *SPINT4* is expressed only in testis and epididymis, and it has been associated with sperm maturation in mice (Penttinen et al. 2003; Clauss et al. 2005). A genome-wide association study identified the Gly73Ser (rs6017667) allele of *SPINT4* as being associated with the multifactorial autoimmune disease Type I diabetes (Todd et al. 2007), which has been previously associated with impairments of male reproductive function in humans (Agbaje et al. 2007; Navarro-Casado et al. 2010).

Considering the distinct selective signatures of *SEMG1*, *WFDC8*, and *SPINT4*, we propose that the selection acting on these genes may be related to innate immune functions in the reproductive tract, with possible consequences for fertility. This hypothesis is easier to reconcile with the geographic restriction of selective signatures and the contribution of different alleles from paralog genes to the overall fitness that could be correlated with host–pathogen interaction and with the pathogen load, which largely differs in type and number across geographic regions (Prugnolle et al. 2005; Barreiro et al. 2008; Coop et al. 2009; Fumagalli et al. 2009, 2011; Pritchard et al. 2010; Seixas et al. 2011). However, because of lack of biological knowledge for some of these genes, the precise form of natural selection driving the departures from neutrality remains unclear.

In summary, we propose that the *WFDC* and *SEMG* loci are under adaptive pressures within the short timescale of modern human evolution. *SEMG1*, *WFDC8*, and *SPINT4* are highlighted as the most likely primary targets of selection in this genomic region. Although the signals found in this locus lead us to hypothesize that immune response to pathogens and fertility drive the selective signatures observed, other unknown biological function(s) of the *WFDC* genes cannot be discarded. Additional studies are needed to address how the molecular evolution of *SEMG1* may alter its biochemical properties and how *WFDC8* and *SPINT4* variants can influence the proteolytic and antimicrobial activity in the human reproductive system.

Materials and Methods

DNA Samples

To study genetic variation in the *WFDC* locus, we resequenced the coding regions of 18 *WFDC* and *SEMG* genes (66 exons total) and a number of intervening noncoding regions (spaced every ~10 kb). In parallel, 47 unrelated, neutrally evolving autosomal regions were polymerase chain reaction (PCR) amplified and sequenced as controls. These regions consist of unlinked, ancient processed pseudogenes expected to evolve neutrally in humans and other primates and were previously used as a proxy for neutral sites (Andrés et al. 2010). See supplementary table S1, Supplementary Material online, for the complete list of loci.

All human samples come from the collection of the International HapMap Project Phase I/II. These included a subset of 21 European (CEU: Utah residents with ancestry from northern and western Europe), 25 African (YRI: Yoruba from Ibadan in Nigeria), and 25 Asian (20 CHB: Han Chinese from Beijing in China and 5 JPT: Japanese from Tokyo in Japan) individuals. See supplementary table S2, Supplementary Material online, for sample identification.

Sequence Generation

Primers for amplification and sequencing of the regions of interest were designed based on the Human Genome Reference Sequence from the March 2006 assembly (v36.1), available at the Genome Browser (http://genome.ucsc.edu/, last accessed January 14, 2013). All samples were PCR amplified and analyzed by bidirectional Sanger sequencing. Further details about PCR and DNA sequencing are available from the authors upon request.

Polymorphic sites were detected with the Phred-Phrap-Consed package (Nickerson et al. 1997). Sites found to have a quality score under 99 were manually curated to minimize sequencing errors. The sequencing data were aligned to the Human RefSeq (hg18), and the ancestral state of each SNP was inferred by comparison with the chimpanzee, orangutan, and macaque genome sequences (Chimpanzee Sequencing and Analysis Consortium 2005; Gibbs et al. 2007; Andrés et al. 2010; http://genome.ucsc.edu/, last accessed January 14, 2013).

Statistical Analysis

The DNA sequence data were analyzed using the classical neutrality tests Tajima's *D*, Fay and Wu's *H*′, Hudson, Kreitman, and Aguade (HKA), and MWU_{high} (Hudson et al. 1987; Tajima 1989; Fay and Wu 2000; Zeng et al. 2005; Nielsen

et al. 2009). Although none of these tests constitutes a formal test of natural selection, they do provide useful metrics for detecting patterns of departure from neutral variation. Tajima's *D* statistic (Tajima 1989) summarizes the polymorphic DNA frequency spectrum; when significantly negative, it is indicative of excess rare variants, consistent with positive selection, purifying selection, or population expansion. When significantly positive, Tajima's *D* identifies a pattern of variation that is consistent with balancing selection or population subdivision, detecting an increased level of common polymorphisms. The Fay and Wu's *H* statistic (Fay and Wu 2000; Zeng et al. 2006) detects an elevated level of high-frequency derived alleles. When significantly negative, Fay and Wu's *H* indicates a signature of a nearly completed selective sweep. MWU_{high} compares the SFS of a region of interest with the SFS from a neutrally evolving region using the MWU statistical test (Nielsen et al. 2009). MWU_{high} is significant only when there is an excess of intermediate-frequency alleles in the locus of interest. Another approach to detect older positive selection is the HKA test (Hudson et al. 1987), which is based on a contrast between polymorphic and fixed differences levels.

MWU was calculated using an in-house C program, whereas Tajima's *D*, Fay and Wu's *H*, and HKA were calculated using the package `libsequence` (Thornton 2003). To control for demographic effects, we assessed the significance of the obtained summary statistics by comparing them to the distributions of statistics from 10,000 neutral demography-corrected coalescent simulations (ms, Hudson 2002), with population recombination estimates predicted from hg18 (http://genome.ucsc.edu/, last accessed January 14, 2013).

Forty-seven neutrally evolving regions (pseudogenes) were sequenced in the CEU, YRI, and Asian (CHB + JPT) populations and analyzed using the same methodology as in Andrés et al. (2010). This was done to further control for the demographic history effects in the studied populations. Among seven demographic models tested (data not shown), the model proposed by Gutenkunst et al. (2009) provided the best fit (goodness of fit) to our control data set. Thus, those were the population demographic parameters that were subsequently used in the neutral coalescent simulations to provide critical values of test statistics.

In addition to the above model-based approach and taking advantage of having sequenced the *WFDC*s and the control regions in the same individuals, we assessed significance of departure from neutrality by contrasting the distribution of test statistics (e.g., Tajima's *D*) generated from the control regions to the observed statistic from each *WFDC* gene. Specifically, we generated an empirical null distribution by calculating these statistics for each of the control regions in each population. We estimated the upper and lower 2.5 percentiles of each distribution and used these thresholds to assess significance of the statistics of each gene.

The levels of population differentiation at the SNP level were calculated with the classical F_{ST} statistic, which describes the proportion of genetic variance attributable to between-population effects (Excoffier 2002). To identify SNPs presenting extreme levels of F_{ST}, the observed F_{ST} at each SNP within the *WFDC* region was compared with the control regions through a locus-by-locus Analysis of Molecular Variance (AMOVA) approach using 20,000 simulations (Arlequin software package; Excoffier et al. 2007).

The potential functional impact of NS SNPs and fixed differences at the protein level was estimated with the PolyPhen-2v HumDiv (Adzhubei et al. 2010) and SIFT (Kumar et al. 2009) algorithms. Although computational predictions are no substitute for molecular studies that identify measureable functional consequences of protein variants, the consistency of SIFT and PolyPhen results combined with the population genetic inferences can be informative.

Haplotype phasing for all samples was inferred separately for the *WFDC*-CEN and *WFDC*-TEL subloci using PHASE2.1 (Stephens et al. 2001; Stephens and Donnelly 2003). Haploview 4.2 (Barrett 2009) used these phased genotype data to calculate LD statistics (r^2 and D') and to identify clusters of high-LD variants (haplotype blocks) (Gabriel et al. 2002). Cladistic (network) relationships among the haplotypes (Bandelt et al. 1999) were infered with Network 4.5.01 software package.

The recent occurrence of an incomplete (or partial) selective sweep is expected to produce a derived haplotype of unusually elevated frequency, and several tests have been devised to detect such events. One of the first of such tests was developed by Hudson et al. (1994), which estimates the probability of finding a subset of haplotypes with a high frequency and low variation, given the total number of segregating sites in the sample. The haplotype test was performed by simulating 10,000 replicates under neutrality with restricted number of segregating sites, incorporating the recombination rate and demographic model previously described (Gutenkunst et al. 2009). To determine statistical significance, the values estimated for the *PI3-SEMG1-SEMG2-SLPI* haplotype were compared against the obtained background neutral distributions. To evaluate the levels of diversity along the haplotype, we calculated values for Tajima's *D*, π, and haplotype diversity using a sliding window (1,000-bp window size and 500-bp increments) and SLIDER online tool (http://genapps.uchicago.edu/slider/index.html, last accessed January 14, 2013).

We also used the DIND test (Barreiro et al. 2009), which considers the ratio of ancestral to derived intrahaplotypic nucleotide diversity ($i\pi_A/i\pi_D$) plotted against the frequency of the derived allele. Specifically, the DIND test was applied to the sequencing data gathered from the *WFDC*-CEN sublocus for each population. A high-frequency-derived allele associated with an elevated $i\pi_A/i\pi_D$ is indicative of an incomplete selective sweep targeting the derived allelic state.

EHH and REHH (Sabeti et al. 2002) were calculated with SWEEP (http://www.broadinstitute.org/mpg/sweep/, last accessed January 14, 2013). The LRH test (Sabeti et al. 2002), performed to assess statistical significance of REHH, included 50 chromosomes simulations under the Gutenkunst demographic model (Gutenkunst et al. 2009; ms, Hudson 2002), for five 500 kb sequence assuming the same population mutation parameter and recombination rate as estimated for the entire *WFDC-CEN* region. Core

haplotypes were set using SWEEP as the longest nonoverlapping cores with no more than three SNPs, with EHH/REHH statistics calculated for 80 kb distance from cores. The significance of EHH/REHH statistics was estimated by comparing values with the null distribution of core haplotypes within the same 5% frequency bin of Ser56.

The T_{MRCA} and neutral parameter θ_{ML} for all the populations were estimated using a maximum likelihood coalescent method implemented in GENETREE version 9 (Griffiths and Tavare 1994). Rare recombinant haplotypes carrying homoplasic mutations were removed from the analysis. We took into account the possibility of selective forces acting on one mutation by estimating the β parameter (Coop and Griffiths 2004). Strictly, the model of Coop and Griffiths constructs a likelihood ratio test on the selection parameter β, contrasting the likelihoods under a null (neutral) model ($\beta = 0$) to that with selection ($\beta \neq 0$), where the neutral model is for a population of constant size. Time, scaled in $2N_e$ generations, was derived from $\theta_{ML} = 4N_e\mu$. The estimate of the mutation rate per gene per generation (μ) was obtained from the average number of nucleotide substitutions per site (D_{xy}) between human and chimpanzee reference sequences, as calculated in DnaSP v.5.1 (Rozas et al. 2003). Time estimates in generations were converted into years using a 25-year generation time. Human/chimpanzee divergence was assumed to have occurred approximately 5.4 Ma (Patterson, Richter, et al. 2006). The likelihood ratio test of Coop and Griffiths was found by simulation to be robust to the impact of demographic change in the parameter range of human population growth. In addition, when we tested for selection in the control genome regions by this test, the neutral null hypothesis was not rejected.

To increase sample size and further test the robustness of our results, we downloaded the corresponding sequenced regions from the 1000 Genomes Project (The 1000 Genomes Project Consortium 2010). We performed a principal component analysis using EIGENSOFT to study the population structure in the *WFDC* locus, using the information of all the SNPs regardless of the LD between them (Patterson, Price, et al. 2006; Price et al. 2006) and calculated summary statistics for every gene in each population using SLIDER. The correlations and independence between the summary statistics of both data sets were determined by Kendall's rank and Spearman's ρ correlations, and χ^2 test.

Supplementary Material

Supplementary tables S1–S7 and figures S1–S7 are available online at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors acknowledge Anh-Dao Nguyen for the help inferring the ancestral allele state of each SNP and Guillaume Laval for making available the scripts for the DIND test. This work was supported in part by the Intramural Research Program of the National Human Genome Research Institute, by the SFRH/BD/45907/2008 fellowship from the Portuguese Foundation for Science and Technology (FCT) to Z.F., by the POPH-QREN-Promotion of scientific employment to Z.F. and S.S., supported by the European Social Fund and national funds of the Portuguese Ministry of Education and Science, and by the Wellcome Trust Centre for Human Genetics (WT097307) to W.W.K. IPATIMUP is an Associated Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT.

References

1000 Genomes Project Consortium. 2010. A map of human genome variation from population-scale sequencing. *Nature* 467: 1061–1073.

Adzhubei IA, Schmidt S, Peshkin L, Ramensky VE, Gerasimova A, Bork P, Kondrashov AS, Sunyaev SR. 2010. A method and server for predicting damaging missense mutations. *Nat Methods.* 7:248–249.

Agbaje IM, Rogers DA, McVicar CM, McClure N, Atkinson AB, Mallidis C, Lewis SE. 2007. Insulin dependant diabetes mellitus: implications for male reproductive function. *Hum Reprod.* 22:1871–1877.

Andrés AM, Dennis MY, Kretzschmar WW, et al. (13 co-authors). 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genet.* 6:e1001157.

Bandelt HJ, Forster P, Rohl A. 1999. Median-joining networks for inferring intraspecific phylogenies. *Mol Biol Evol.* 16:37–48.

Barreiro LB, Ben-Ali M, Quach H, et al. (18 co-authors). 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. *PLoS Genet.* 5:e1000562.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L. 2008. Natural selection has driven population differentiation in modern humans. *Nat Genet.* 40:340–345.

Barrett JC. 2009. Haploview: visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc.* 2009:pdb.ip71.

Benjamini Y, Hochberg Y. 1995. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J R Stat Soc Ser B.* 57:289–300.

Bingle CD, Vyakarnam A. 2008. Novel innate immune functions of the whey acidic protein family. *Trends Immunol.* 29:444–453.

Bouchard D, Morisset D, Bourbonnais Y, Tremblay GM. 2006. Proteins with whey-acidic-protein motifs and cancer. *Lancet Oncol.* 7: 167–174.

Bourgeon F, Evrard B, Brillard-Bourdet M, Colleu D, Jegou B, Pineau C. 2004. Involvement of semenogelin-derived peptides in the antibacterial activity of human seminal plasma. *Biol Reprod.* 70:768–774.

Carnahan SJ, Jensen-Seaman MI. 2008. Hominoid seminal protein evolution and ancestral mating behavior. *Am J Primatol.* 70:939–948.

Chowdhury MA, Kuivaniemi H, Romero R, Edwin S, Chaiworapongsa T, Tromp G. 2006. Identification of novel functional sequence variants in the gene for peptidase inhibitor 3. *BMC Med Genet.* 7:49.

Clauss A, Lilja H, Lundwall A. 2005. The evolution of a genetic locus encoding small serine proteinase inhibitors. *Biochem Biophys Res Commun.* 333:383–389.

Clauss A, Persson M, Lilja H, Lundwall.Å. 2011. Three genes expressing Kunitz domains in the epididymis are related to genes of WFDC-type protease inhibitors and semen coagulum proteins in spite of lacking similarity between their protein products. *BMC Biochem.* 12:55.

Chimpanzee Sequencing and Analysis Consortium C. 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437:69–87.

Coop G, Griffiths RC. 2004. Ancestral inference on gene trees under selection. *Theor Popul Biol.* 66:219–232.

Coop G, Pickrell JK, Novembre J, Kudaravalli S, Li J, Absher D, Myers RM, Cavalli-Sforza LL, Feldman MW, Pritchard JK. 2009. The role of geography in human adaptation. *PLoS Genet.* 5:e1000500.

de Lamirande E. 2007. Semenogelin, the main protein of the human semen coagulum, regulates sperm function. *Sem Thromb Hemost.* 33:60–68.
Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT. 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nat Genet.* 36:1326–1329.
Drmanac R, Sparks AB, Callow MJ, et al. (65 co-authors). 2010. Human genome sequencing using unchained base reads on self-assembling DNA nanoarrays. *Science* 327:78–81.
Edstrom AM, Malm J, Frohm B, Martellini JA, Giwercman A, Morgelin M, Cole AM, Sorensen OE. 2008. The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins. *J Immunol.* 181: 3413–3421.
Excoffier L. 2002. Human demographic history: refining the recent African origin model. *Curr Opin Genet Dev.* 12:675–682.
Excoffier L, Laval G, Schneider S. 2007. Arlequin (version 3.0): an integrated software package for population genetics data analysis. *Evol Bioinform Online.* 1:47–50.
Fay JC, Wu C-I. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.
Ferreira Z, Hurle B, Rocha J, Seixas S. 2011. Differing evolutionary histories of WFDC8 (short-term balancing) in Europeans and SPINT4 (incomplete selective sweep) in Africans. *Mol Biol Evol.* 28: 2811–2822.
Fornarino S, Laval G, Barreiro LB, Manry J, Vasseur E, Quintana-Murci L. 2011. Evolution of the TIR domain-containing adaptors in humans: swinging between constraint and adaptation. *Mol Biol Evol.* 28: 3087–3097.
Fu YX. 1996. New statistical tests of neutrality for DNA samples from a population. *Genetics* 143:557–570.
Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M. 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. *Genome Res.* 19:199–212.
Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R. 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. *PLoS Genet.* 7:e1002355.
Gabriel SB, Schaffner SF, Nguyen H, et al. (18 co-authors). 2002. The structure of haplotype blocks in the human genome. *Science* 296: 2225–2229.
Gibbs RA, Rogers J, Katze MG, et al. (177 co-authors). 2007. Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316:222–234.
Griffiths RC, Tavare S. 1994. Sampling theory for neutral alleles in a varying environment. *Philos Trans R Soc Lond B Biol Sci.* 344: 403–410.
Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD. 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. *PLoS Genet.* 5: e1000695.
Hancock AM, Clark VJ, Qian Y, Di Rienzo A. 2011. Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. *Mol Biol Evol.* 28:601–614.
Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M. 2011. Classic selective sweeps were rare in recent human evolution. *Science* 331:920–924.
Hudson RR. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18:337–338.
Hudson RR, Bailey K, Skarecky D, Kwiatowski J, Ayala FJ. 1994. Evidence for positive selection in the superoxide dismutase (Sod) region of *Drosophila melanogaster*. *Genetics* 136:1329–1340.
Hudson RR, Kreitman M, Aguadé M. 1987. A test of neutral molecular evolution based on nucleotide data. *Genetics* 116:153–159.
Hurle B, Swanson W, Program NCS, Green ED. 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. *Genome Res.* 17:276–286.
Jensen-Seaman MI, Li WH. 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *J Mol Evol.* 57: 261–270.
Kingan SB, Tatar M, Rand DM. 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *J Mol Evol.* 57:159–169.
Kumar P, Henikoff S, Ng PC. 2009. Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat Protoc.* 4:1073–1082.
Lundwall A. 2007. A locus on chromosome 20 encompassing genes that are highly expressed in the epididymis. *Asian J Androl.* 9: 540–544.
Lundwall A, Bjartell A, Olsson AY, Malm J. 2002. Semenogelin I and II, the predominant human seminal plasma proteins, are also expressed in non-genital tissues. *Mol Hum Reprod.* 8:805–810.
Lundwall A, Clauss A. 2011. Genes encoding WFDC- and Kunitz-type protease inhibitor domains: are they related? *Biochem Soc Trans.* 39: 1398–1402.
Martellini JA, Cole AL, Venkataraman N, Quinn GA, Svoboda P, Gangrade BK, Pohl J, Sorensen OE, Cole AM. 2009. Cationic polypeptides contribute to the anti-HIV-1 activity of human seminal plasma. *FASEB J.* 23:3609–3618.
McKiernan PJ, McElvaney NG, Greene CM. 2011. SLPI and inflammatory lung disease in females. *Biochem Soc Trans.* 39:1421–1426.
McNeely TB, Shugars DC, Rosendahl M, Tucker C, Eisenberg SP, Wahl SM. 1997. Inhibition of human immunodeficiency virus type 1 infectivity by secretory leukocyte protease inhibitor occurs prior to viral reverse transcription. *Blood* 90:1141–1149.
Navarro-Casado L, Juncos-Tobarra MA, Chafer-Rudilla M, Iniguez-de Onzono L, Blazquez-Cabrera JA, Miralles-Garcia JM. 2010. Effect of experimental diabetes and STZ on male fertility capacity: study in rats. *J Androl.* 31:584–592.
Nickerson DA, Tobe VO, Taylor SL. 1997. PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res.* 25:2745–2751.
Nielsen R, Hubisz MJ, Hellmann I, et al. (13 co-authors). 2009. Darwinian and demographic forces affecting human protein coding genes. *Genome Res.* 19:838–849.
Patterson N, Price AL, Reich D. 2006. Population structure and eigenanalysis. *PLoS Genet.* 2:e190.
Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D. 2006. Genetic evidence for complex speciation of humans and chimpanzees. *Nature* 441:1103–1108.
Penttinen J, Pujianto DA, Sipila P, Huhtaniemi I, Poutanen M. 2003. Discovery in silico and characterization in vitro of novel genes exclusively expressed in the mouse epididymis. *Mol Endocrinol.* 17: 2138–2151.
Peter A, Lilja H, Lundwall A, Malm J. 1998. Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase. *Eur J Biochem.* 252:216–221.
Price AL, Patterson NJ, Plenge RM, Weinblatt ME, Shadick NA, Reich D. 2006. Principal components analysis corrects for stratification in genome-wide association studies. *Nat Genet.* 38:904–909.
Pritchard JK, Pickrell JK, Coop G. 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. *Curr Biol.* 20:208–215.
Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F. 2005. Pathogen-driven selection and worldwide HLA class I diversity. *Curr Biol.* 15:1022–1027.
Przeworski M, Coop G, Wall JD. 2005. The signature of positive selection on standing genetic variation. *Evolution* 59(11):2312–2323.
Rajesh A, Madhubabu G, Yenugu S. 2011. Identification and characterization of Wfdc gene expression in the male reproductive tract of the rat. *Mol Reprod Dev.* 78:633–641.
Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD. 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. *Mol Biol Evol.* 25:207–219.
Robert M, Gagnon C. 1999. Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein. *Cell Mol Life Sci.* 55:944–960.

Robert M, Gibbs BF, Jacobson E, Gagnon C. 1997. Characterization of prostate-specific antigen proteolytic activity on its major physiological substrate, the sperm motility inhibitor precursor/semenogelin I. *Biochemistry* 36:3811–3819.

Rozas J, Sanchez-DelBarrio JC, Messeguer X, Rozas R. 2003. DnaSP, DNA polymorphism analyses by the coalescent and other methods. *Bioinformatics* 19:2496–2497.

Sabeti PC, Reich DE, Higgins JM, et al. (17 co-authors). 2002. Detecting recent positive selection in the human genome from haplotype structure. *Nature* 419:832–837.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D. 2005. Calibrating a coalescent simulation of human genome sequence variation. *Genome Res.* 15:1576–1583.

Seixas S, Ivanova N, Ferreira Z, Rocha J, Victor BL. 2011. Loss and Gain of Function in SERPINB11: an example of a gene under selection on standing variation, with implications for host-pathogen interactions. *PLoS One* 7(2):e32518.

Stephens M, Donnelly P. 2003. A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet.* 73:1162–1169.

Stephens M, Smith NJ, Donnelly P. 2001. A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet.* 68:978–989.

Storey JD. 2002. A direct approach to false discovery rates. *J R Stat Soc Ser B.* 64:479–498.

Storey JD, Taylor JE, Siegmund D. 2004. Strong control, conservative point estimation, and simultaneous conservative consistency of false discovery rates: a unified approach. *J R Stat Soc Ser B.* 66: 187–205.

Storey JD, Tibshirani R. 2003. Statistical significance for genome-wide experiments. *Proc Natl Acad Sci U S A.* 100:9440–9445.

Tajima F. 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123:585–595.

Thimon V, Calvo E, Koukoui O, Legare C, Sullivan R. 2008. Effects of vasectomy on gene expression profiling along the human epididymis. *Biol Reprod.* 79:262–273.

Thornton K. 2003. Libsequence: a C++ class library for evolutionary genetic analysis. *Bioinformatics* 19:2325–2327.

Todd JA, Walker NM, Cooper JD, et al. (40 co-authors). 2007. Robust associations of four new chromosome regions from genome-wide analyses of type 1 diabetes. *Nat Genet.* 39:857–864.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A. 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. *Proc Natl Acad Sci U S A.* 102:18508–18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK. 2006. A map of recent positive selection in the human genome. *PLoS Biol.* 4:e72.

Wang Z, Widgren EE, Sivashanmugam P, O'Rand MG, Richardson RT. 2005. Association of eppin with semenogelin on human spermatozoa. *Biol Reprod.* 72:1064–1070.

Watterson GA. 1975. On the number of segregating sites in genetical models without recombination. *Theor Popul Biol.* 7:256–276.

Weldon S, McGarry N, Taggar CC, McElvaney NG. 2007. The role of secretory leucoprotease inhibitor in the resolution of inflammatory responses. *Biochem Soc Trans.* 35:273–276.

Weldon S, Taggart CC. 2007. Innate host defense functions of secretory leucoprotease inhibitor. *Exp Lung Res.* 33:485–491.

Whitlock MC. 2005. Combining probability from independent tests: the weighted Z-method is superior to Fisher's approach. *J Evol Biol.* 18: 1368–1373.

Williams SE, Brown TI, Roghanian A, Sallenave JM. 2006. SLPI and elafin: one glove, many fingers. *Clin Sci.* 110:21–35.

Yenugu S, Richardson RT, Sivashanmugam P, Wang Z, O'Rand M G, French FS, Hall SH. 2004. Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif. *Biol Reprod.* 71:1484–1490.

Zeng K, Fu YX, Shi S, Wu CI. 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174: 1431–1439.

Zhao H, Lee WH, Shen JH, Li H, Zhang Y. 2008. Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma. *Peptides* 29:505–511.

3.2. Genetic diversity and evolution of *WFDC*s in chimpanzees

3.2.1. Sequence diversity of *Pan troglodytes* subspecies and the impact of *WFDC6* selective constraints in reproductive immunity.

Submitted

Classification: Biological Sciences - Genetics

Sequence diversity of *Pan troglodytes* subspecies and the impact of *WFDC6* selective constraints in reproductive immunity

Research Article

Zélia Ferreira[1,2,3 *], Belen Hurle[1 *], Aida M. Andrés[4], Warren Kretzschmar[5], Jim Mullikin[6], Praveen Cherukuri[6], Pedro Cruz[6], Mary Katherine Gonder[7], Anne Stone[8], Sarah Tishkoff[9], Willie Swanson[10], NISC Comparative Sequencing Program[1,6], Eric D. Green[1], Andrew G. Clark[11], and Susana Seixas[2].

1 - National Human Genome Research Institute, National Institutes of Health, 9000 Rockville Pike, Bethesda, MD 20892-2152, United States of America.

2 - Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto 4200-465, Portugal;

3 - Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto 4099-002, Portugal;

4 - Department of Evolutionary Genetics – Max Planck Institute for Evolutionary Anthropology, Leipzig 04103, Germany;

5 - Genomic Medicine and Statistics, Wellcome Trust Centre for Human Genetics, University of Oxford, Oxford, United Kingdom;

6 - NIH Intramural Sequencing Center (NISC), National Human Genome Research Institute, National Institutes of Health, Rockville, Maryland 20852, USA;

7 - Department of Biological Sciences University at Albany, State University of New York, Albany NY;

8 - School of Human Evolution and Social Change, Arizona State University, Tempe, AZ 85287, USA;

9 - Departments of Genetics and Biology, University of Pennsylvania, Philadelphia, PA;

10- Department of Genome Sciences, University of Washington, Seattle, Washington 98195, USA;

11 – Department of Biology of Molecular Biology and Genetics, Cornell University, Ithaca, New York 11850, USA;

* These authors contributed equally to this work

Corresponding author: Zélia Ferreira

Graduate Student, Faculty of Sciences of the University of Porto, Institute of Molecular Pathology and Immunology of the University of Porto and National Human Genome Research Institute, National Institutes of Health.

Rua Dr. Roberto Frias s/n

4200-465 Porto, Portugal

zferreira@ipatimup.pt - email

+351225570700 – phone

keywords: WFDC, natural selection, innate immunity, serine protease inhibitor, reproduction, chimpanzee

Running head: Chimpanzee diversity and conservation of *WFDC6*

Abstract

Recent efforts have attempted to describe the population structure of common chimpanzee, focusing on four subspecies: *Pan troglodytes verus*, *P. t. ellioti*, *P. t. troglodytes*, and *P. t. schweinfurthii*. However, few studies have pursued the effects of natural selection in shaping their response to pathogens and reproduction. Whey acidic protein (WAP) four-disulfide core domain (*WFDC*) genes and neighboring semenogelins (*SEMG*) genes encode proteins with combined roles in immunity and fertility. They display a strikingly high rate of amino acid replacement (d_N/d_S), indicative of adaptive pressures during primate evolution. In human populations, three signals of selection at the *WFDC* locus were described, possibly influencing the proteolytic profile and antimicrobial activities of the male reproductive tract. To evaluate the patterns of genomic variation and selection at the WFDC locus in chimpanzees, we sequenced 17 *WFDC* genes and 47 autosomal pseudogenes in 68 chimpanzees (15 *P. t. troglodytes*, 22 *P. t. verus*, and 31 *P. t. ellioti*). The genomic diversity and substructure of the common chimpanzee subspecies was revisited based on our data. We found a clear differentiation of *P. t. verus* from *P. t. troglodytes* and *P. t. ellioti* subspecies; further, at the *WFDC* locus, we identified a signature of strong selective constraints common to the three subspecies in *WFDC6* - a recent paralog of the epididymal protease inhibitor *EPPIN*. Overall, chimpanzees and humans do not display similar footprints of selection across the *WFDC* locus, possibly due to different selective pressures between the two species related to immune response and reproductive biology.

Introduction

Common chimpanzees and bonobos are different species of the *Pan* genus (*P. troglodytes* and *P. paniscus*, respectively), separated by the geographical barrier of the Congo River. Common chimpanzees are further divided into subspecies across tropical Africa (1, 2). The issue of genomic diversity and substructure among the different chimpanzee subspecies is controversial and of great interest. Briefly, *P. troglodytes* was traditionally divided in three subspecies: *P. t. verus*, located in western Africa occupying the Upper Guinea region; *P. t. troglodytes* extending throughout central Africa; and *P. t. schweinfurthii* living in eastern Africa. Later, the analysis of mitochondrial DNA (mtDNA) variation led to the proposal of a fourth chimpanzee subspecies, *P. t. ellioti* (also known as *P. t. vellerosus)*, occurring in the Gulf of Guinea (Nigeria and Cameroon) in a region limited by the Niger and Sanaga rivers (Fig S1) (2-5). Recent studies support the differentiation of *P. t. ellioti* from *P. t. troglodytes* using ancestry-informative markers, enabling the identification of the four subspecies (2, 6). *P.t. verus* branched from the last common chimpanzee ancestor ~0.46 million years ago (mya), and *P. t. ellioti* diverged from *P. t. troglodytes* and *P. t. schweinfurthii* ~0.32 mya. Even though occasional hybridization occurs between *P. t. ellioti* and *P. t. troglodytes* in the wild, these subspecies remain as major genetic isolates (2). Studies regarding chimpanzee simian immunodeficiency virus (SIVcpz) also support these findings, given that only *P. t. troglodytes* and *P. t. schweinfurthii* are infected in the wild (>30% prevalence), and *P. t. ellioti*, only get infected when kept in captivity with *P. t. troglodytes* (7, 8). SIVcpz is one of many infectious agents transferred to humans from chimpanzees originating the human immunodeficiency virus (HIV) (9). Therefore, a better characterization of episodes of natural selection in chimpanzees may provide ways to further understand susceptibility to pathogens in hominoids, and to improve the conservation of wild chimpanzees.

Ecological changes in the natural habitat of *P. troglodytes* have modeled immune response to pathogens and reproductive-driven phenotypes (10). A genomic locus that is involved in both immune response and reproduction is the whey acidic protein (WAP) four-disulfide core domain (*WFDC*) locus (Fig 1). *WFDC* genes (17 in total) encode small serine protease inhibitors with functions of regulating endogenous proteases (11, 12). Neighboring genes, semenogelin 1 and 2 (*SEMG1* and *SEMG2*) encode the main proteins of the seminal coagulum (12-14). *WFDC* and *SEMG* genes evolved from the

same common ancestor and maintain some similar functions involving antimicrobial, immune, and male reproduction activities (15-17). Well-characterized genes at the *WFDC* locus include peptidase inhibitor 3 (PI3; also known as elafin) and secretory leucocyte proteinase inhibitor (SLPI), both pleiotropic molecules synthesized at mucosal surfaces that play a role in the surveillance against microbial and viral infections, including HIV-1 (18). This locus also includes the epididymal protease inhibitor *EPPIN* (also known as *SPINLW1*), which coats the surface of human spermatozoa, binds to SEMG1, and modulates the activity of prostate-specific antigen (PSA) altogether providing antimicrobial protection for spermatozoa (19-21). SEMG1 and SEMG2 play critical roles in semen clotting and in antimicrobial and antiviral protection for spermatozoa in the female reproductive tract (20, 22). *WFDC* and *SEMG* genes were shown to be targets of adaptive evolution in primates, where SEMGs d_N/d_S values were positively correlated with female promiscuity. Specifically, in monoandrous primates in which females mate with a single male (*e.g*. humans, gorillas, and gibbons), the ejaculate is gelatinous in contrary to polyandrous primates, in which females mate with multiple partners per ovulatory period (*e.g*. chimpanzees and macaques), the ejaculate forms a rigid copulatory plug that prevents the insemination of females by competing males (23, 24). In chimpanzees, the copulatory plug formation is associated with a SEMG1 length expansion causing an increase in protein crosslinking (25).

In human populations, the *WFDC* locus presents complex selective signals, including recent balancing selection on *WFDC8* in Europeans, and positive selection in *SEMG1* in Asians (26). In order to evaluate the patterns of genomic variation and selection at the *WFDC* locus in chimpanzees, we sequenced 18 *WFDC* genes and 47 control regions in 68 common chimpanzees from the subspecies *P. t. troglodytes*, *P. t. verus,* and *P. t. ellioti*. Overall, we generated a total of ~13 Mb of high-quality sequence data, identified 1268 Single Nucleotide Polymorphisms (SNPs), and calculated summary statistics of population variation for 71 loci. The controversial issue of genomic diversity among the different chimpanzee subspecies was also revisited based on our generated data. We found a clear differentiation of *P. t. verus* from *P. t. troglodytes* and *P. t. ellioti* subspecies and in general, for *WFDC* genes, we did not find departures from diversity levels observed in neutral evolving regions. Notwithstanding, we identified a signature of strong selective constraints common to the three studied subspecies and centered in the *EPPIN-like* gene, *WFDC6*. In several primate species,

WFDC6 has lost the ability to inhibit PSA and in others it appears to have accumulated different deleterious mutations. Conversely, in chimpanzees and humans, the seven disulfide bridges, known to confer antimicrobial properties to *WFDC* genes, are preserved in *WFDC6*. The fact that chimpanzees have a polyandrous mating system and as a promiscuous species they are particularly likely to be exposed to sexually transmitted pathogens, lead us to propose that strong conservation of *WFDC6* function has been necessary in chimpanzees due to its crucial role in innate immunity of the reproductive tract.

Results

Polymorphism Levels and Population Structure

We generated sequence data from 68 chimpanzees for all *WFDC* exons distributed across 54 amplicons and 47 neutrally evolving control regions, for a total of 13 Mb (Table S1). This resulted in the identification of 419 SNPs in the control regions and 849 SNPs in the *WFDC* locus. Despite the high-quality sequence data, the patterns of variation in *SEMG1* among individuals cannot be analyzed due to a variable and highly repetitive region (9 to 13 copies) in this gene (25). The modular nature of *SEMG1* precluded a consistent and unambiguous sequence alignment; therefore, the SNPs located in this repetitive region were removed from the analysis for quality purposes. A total of 766 fixed differences were identified in comparison with the human genome reference. Twenty-five indels (insertions/deletions) were found, 24 of which were located in introns, UTRs and intergenic regions. One remaining indel located in *WFDC6* was identified in a single chromosome (f=0.02). Indels were excluded from all analyses, due to their distinct mutation rate and to the low overlap with functional regions, which are unlikely to affect protein function or expression. Additionally, from the 456 human-chimpanzee fixed differences located in the *WFDC* locus only 19 were within coding regions and chimpanzee-specific. Of these, 17 were non-synonymous (NSyn) and most of them were classified as benign by Polyphen v2 and SIFT (Table S4b).

To assess the selective pressures shaping the diversity in *WFDC* genes, we first characterized the levels of differentiation level between subspecies at the control regions and *WFDC* locus, by using F_{ST} (27) and principal component analysis (PCA) (28). The pairwise F_{ST} values show lower differentiation between *P. t. troglodytes* and *P. t. ellioti* (0.15) than each of them compared with *P. t. verus* (0.3761 and 0.3255, respectively). The differentiation levels of *WFDC* genes are not significantly higher than in the control regions, suggesting the differentiation in the *WFDC*s has been mainly shaped by the demographic history of each subspecies (Fig S3A). The first two principal components (PC1 and PC2) separate *P. t. verus* from the other two subspecies, and the third principal component appears to not separate completely *P. t. troglodytes* from *P. t. ellioti* (Fig S4). Contrary to previous studies we could not separate the three subspecies using PC analysis (2, 6). We further examined the

shared ancestry levels of the individuals, by performing a Bayesian model-based clustering approach available in the STRUCTURE software (29, 30). The analyses were performed blinded to a population label and two groups (K=2) were recovered, despite the value for maximum likelihood (KMAX) being 3 (Fig S5). The subspecies *P. t. troglodytes* and *P. t. ellioti* could not be confidently distinguished.

To characterize the within subspecies variation, we analyzed the folded site frequency spectrum (SFS) for all SNPs highlighting synonymous (Syn) and NSyn sites among *WFDC* genes (Fig 2A and B) and in the control regions (Fig S2A). Despite the higher number of NSyn sites in the *WFDC* genes, these are maintained at low frequencies in the overall species, consistent with the predicted mildly deleterious effects (Table S4). For each *WFDC* gene, we calculated summary statistics such as nucleotide diversity (π), Tajima's *D* (31), Fu and Li's *D* (32), Fay and Wu's *H* (33, 34), and HKA (35) (Table S3).

Analyses of SFS and summary statistics show *P. t. troglodytes* as the subspecies with the highest nucleotide diversity levels, and *P. t. verus* as the most homogeneous subspecies (Fig S2B; Table S3). Both *P. t. troglodytes* and *P. t. ellioti* Tajima's *D* values are skewed toward negative values, mostly due to their large effective population size and a population expansion that is estimated to have occurred around 50,000 years ago (kya) (36). Nonetheless, *P. t. verus* Tajima's *D* values are less negative than the other two subspecies, which is likely to result from an extreme decrease in population size and genetic drift (36-38). Overall, the analyses of the summary statistics of *WFDC* genes do not show widespread significant departure from neutrality, showing only mildly negative or positive values.

Selection Tests

To determine if specific *WFDC* genes have been under selective pressures in one or all chimpanzee subspecies, we started by comparing the summary statistics for each *WFDC* gene with the empirical distribution of π and Tajima's *D* in the control regions (Fig S7). Only *WFDC6* and *EPPIN* show unusual patterns in *P. t. troglodytes* (Fig S7). Although in this subspecies the allele frequency spectrum is generally skewed toward rare alleles, the Tajima's *D* values of *WFDC6* (-2.073; *P*-value = 1^{-4}) and *EPPIN* (-1.811; *P*-value = 0.025) present the lowest values of control and *WFDC* regions.

WFDC6 also presented a low Tajima's *D* value (-2.1039) and significant HKA (*P* = 0.013) when combining all individuals sequenced. The other *WFDC* genes did not show significant *P*-values pointing to a neutral evolution based on subspecies genetic diversity (Table S3). To confirm *WFDC6* and *EPPIN* departure from neutrality, we performed 10^5 coalescent simulations under three different demographic scenarios: Constant model, population expansion dated 50 kya, and a best-fit model for *P. t. troglodytes* (36). *WFDC6* presents a significantly negative Tajima's *D* value compared to all models, but *EPPIN* loses significance when taking into account the demographic history of *P. t. troglodytes* [(36); Table S5].

The hypothesis of a recent positive selection was excluded due to absence of LD blocks and homogeneous haplotypes in *P. t. troglodytes* (Fig S6), which prevent long-range haplotype tests from being calculated. To address the hypothesis of an older selective sweep we performed the McDonald-Kreitman test (MKT), which did not show departures from neutrality in either *WFDC6* (*P*-value = 0.74) or *EPPIN* (*P*-value = 0.75). Notwithstanding, population differentiation (F_{ST} statistic) in the *WFDC6-EPPIN* region is the lowest in the *WFDC* locus (Fig S3B), and the networks built to assess the *WFDC6* and *EPPIN* haplotype structure show that *WFDC6* has a star-shaped genealogy shared among all subspecies (Fig 3; Fig S8). The findings show that in *WFDC6* all NSyn variants are maintained a very low frequencies and that the fixed difference K79E, predicted to alter protein function in EPPIN, is also present in bonobos (>1Mya). This suggests strong purifying selection as the likeliest cause for *WFDC6* patterns of diversity.

To determine the levels of selective constraints operating at *WFDC6* and *EPPIN*, we aligned the publicly available sequences of both genes for 8 primate species (chimpanzee, human, gorilla, orangutan, gibbon, rhesus monkey, baboon and marmoset). The alignment shows that *EPPIN* has been conserved in all the species. *WFDC6* has released constraints with signals of pseudogenization in orangutan, rhesus and baboon and appears to be absent in marmoset (Fig S10). Evidences for *WFDC6* pseudogenization includes a premature stop codon (W86X) in orangutan, a very early stop codon (S4X) and a five amino acid deletion (28 to 32) in rhesus monkey, and a frameshift mutation (T99fs139X) shared between rhesus and baboon. Also striking is the loss of two disulfide bridges in the Kunitz domain for all the primates, the species-specific loss of one disulfide bridge in the WAP domain from

gorilla and gibbon; and the loss of the SEMG1-binding residue in gibbon. Furthermore, the active site conferring PSA inhibitory activity to EPPIN, in WFDC6 was modified from a leucine to a tryptophan (residue 87) in most primates, and to a stop codon in orangutan (Fig S9).

We calculated d_N/d_S (ω; d_S – synonymous substitution rate, and d_N - non-synonymous substitution rate) ratios for the paralogs *WFDC6* and *EPPIN*, under alternative models of gene evolution, for the entire sequence dataset after the exclusion of *WFDC6* pseudogenes (orangutan, rhesus and baboon sequences). In cases where no selection is operating, ω should be equal to one, below one when purifying selection is acting to preserve protein sequence, and significantly exceed one when positive selection is acting to drive divergence of protein sequence. We estimated a single ω value for the entire phylogeny (one-ratio), in which we assumed no differentiation in *WFDC6* and *EPPIN* selective pressures. The observed value (ω = 0.4739) is lower than one suggesting an overall conservation of *WFDC6* and *EPPIN* (Table S6) (39). As these two proteins are very similar, we examined if the two paralogs have been subject to different selective constraints, and applied the two-ratio model, considering the branches that correspond to *WFDC6* and to *EPPIN* clades. The ω value for *WFDC6* was close to 1 (ω_{WFDC6} = 0.8846) and almost two times higher than EPPIN (ω_{EPPIN} = 0.4738) but they did not differ significantly from the one-ratio model (-2ΔI = 1.04; *P*-value = 0.35). To determine if *WFDC6* was under different selective constraints in chimpanzees, we performed two more tests: a three-ratio model, where we define the human and chimpanzee and their ancestral as one clade, and another three-ratio model, where we define only the chimpanzee *WFDC6* as an independent clade. Although our results suggest that *WFDC6* might have experienced very different selective pressures, as shown by the human-chimpanzee $\omega_{ancWFDC6}$ = 1.1656, and by the chimpanzee ω_{WFDC6} = 0.5886, none of the new tests indicate a significant departure from the two-ratio model (Table S6). However, the lack of significance might be related to the small dataset tested and to the short sequence length (only 131 codons).

Discussion

Here, we studied the sequence diversity at the *WFDC* locus and 47 neutrally evolving regions chosen to control for demographic effects in three *Pan troglodytes*

subspecies, *P. t. troglodytes*, *P. t. ellioti*, and *P. t. verus*. In our dataset, we inferred the strength of the selective pressures acting in *WFDC* locus after retrieving the structural and geographical differentiation of the three chimpanzee subspecies. This analysis shows that *P. t. verus* is the least diverse subspecies and a clear defined genetic entity, while *P. t. ellioti* and *P. t. troglodytes* subspecies are more diverse and hardly discriminated even with a set of 1268 autosomal SNPs. In the *WFDC* locus, we pinpointed a single selective signal, which has a high degree of inter-population homogeneity and identifies *WFDC6* as a gene under purifying selection in chimpanzees. We hypothesize that these selective constraints were driven by a response to sexually transmitted pathogens, as *Pan troglodytes* is a promiscuous species and gets infected by a plethora of infectious diseases in the wild.

The controversial issue of genomic diversity and substructure of the common chimpanzee subspecies was revisited based on the variation data collected for this study. Most studies addressing the genetic variation in *Pan troglodytes* focus on sequencing mtDNA and study microsatellites, sometimes combined with ascertained autosomal markers, to infer demographic history (2, 6). The contribution of our data to the complex question of chimpanzee demographic history is modest. This can be due to the lack of markers in the 47 control regions (419 SNPs) to discriminate *P. t. ellioti* and, *P. t. troglodytes*; and to the fact that the *WFDC* locus may be not the best representative of the overall genome diversity in chimpanzees. The *WFDC* region has been pinpointed as being under adaptive evolution in previous studies (40-42), potentially reducing genetic distance estimates used to infer demographic history.

Nevertheless, we detected a consistent differentiation of *P. t. verus* from the other two subspecies, and confirmed that the nucleotide diversity of *P. t. verus* is more similar to humans. Generally, *P. t. troglodytes* is the subspecies with the highest levels of diversity, and *P. t. ellioti* does not differ significantly (2, 6). As expected, a larger effective population size in *P. t. troglodytes* explains the higher levels of diversity, and the negative Tajima's *D* trend is consistent with a history of population expansion (36, 38, 43-45). These trends are also clear in *P. t. ellioti*, suggesting a similar population history and relatedness (6, 8). Despite *P. t. verus* and humans having similar levels of nucleotide diversity in *WFDC*s, the patterns of LD are considerably different. Some progress has been made regarding recombination rates in chimpanzees, and the only genome-wide available catalog, PanMap, was calculated from 10 individuals of the *P. t. verus* subspecies. This rate in the *WFDC* region is in the same of order of magnitude

as in humans (~ 0.2 cM/Mb) but only a few and narrow haplotype blocks were identified in *P. t. verus* subspecies.

The signatures of selection identified in the human *WFDC* locus are mainly associated with homogeneous long-range haplotypes and variants located at *SEMG1* (Asians), *WFDC8* (Europeans), and *SPINT4* (Africans) indicating a recent increase in the frequency by selection and not by demographic events (26). In chimpanzees, we assessed the signatures of positive selection by circumventing the effects of demography and comparing the *WFDC* genes with the empirical distribution built from 47 neutrally evolving regions of the chimpanzee genome, as we did previously in humans (26). Even though we could not detect LD blocks or extended haplotypes in our sequenced data, we find lower levels of nucleotide diversity in *WFDC6* and *EPPIN* genes while compared with other chimpanzee loci. Both genes present levels of nucleotide diversity that are similar to humans, in all subspecies. They also present significantly negative Tajima's *D* values in the total number of individuals analyzed and in *P. t. troglodytes*, using either coalescent simulations or empirical comparisons. The significantly negative HKA *P*-value obtained for the total sample set together with the low subspecies differentiation indicated by F_{ST} and the low frequencies of NSyn variants is suggestive of a signature of an old event of purifying selection in *WFDC6* and *EPPIN* in *Pan troglodytes*.

To our knowledge, no experimental studies were performed to determine WFDC6 biological functions. However, *WFDC6* is considered a recent paralog of *EPPIN*, with 71% of sequence similarity and a very similar pattern of expression in male reproductive tissues (15). At the protein level, WFDC6 and EPPIN share the same domains (WAP and Kunitz), and thus we might assume that WFDC6 has similar functions to EPPIN, including binding to SEMG1. EPPIN is known to protect SEMG1 from premature cleavage by its natural protease, PSA, a protease inhibitor activity conferred by L87 residue (P1 reactive site) located in the Kunitz domain. Other recognized roles of EPPIN are its antimicrobial and antiviral activities, providing protection of the spermatozoa (15). Due to its important functions in reproduction in primates, it is not unexpected that some level of purifying selection is acting on EPPIN to prevent NSyn mutations from altering its important biological functions. Even in primates experiencing lower levels of post-copulatory selection and lower semen

coagulum thickness like gorilla (41), it seems that the role of EPPIN in modulating the cleavage of SEMG1 is not affected. However, *WFDC6*, which shows the strongest signature of purifying selection in chimpanzees does not share the same leucine residue at the reactive site, instead it has in position 87 a tryptophan (W). It is also noticeable that the majority of the replacements seen in WFDC6 include cysteines from the Kunitz domain, which in EPPIN are engaged in disulfide bonds. Therefore, we hypothesize that the serine-protease activity of WFDC6 would be impaired or targeted to a different protease other than PSA. On the other hand, the WAP domain has a highly conserved amino acidic composition between both genes, and the maintenance of the disulfide bridges suggest the antimicrobial properties of this domain will be maintained. The WAP motif-containing proteins are functionally diverse, including serine-protease inhibition (11), immunity (46), antiviral (18, 47), and tumorogenesis (48).

Disease transmission during mating provides a connection between reproduction and immunity, where sexually transmitted diseases (STDs) can affect fitness of individuals by imposing different selective pressures on their hosts. Previous studies found a positive correlation between levels of leukocytes (indicator of immunocompetence) and several proxies of female sexual promiscuity among species of primates with different mating systems (49-51). The lack of associations with several other social, ecological and life-history variables led to the hypothesis that increased levels of transmission of STDs in promiscuous species have resulted in the evolution of a greater investment in immune response (52, 53). Chimpanzees are classified as one of the most promiscuous primate species, where previous signals of rapid evolution of sperm proteins (SEMG1 and SEMG2) were found (25, 40, 41). Instead, humans, gorillas and gibbons are not promiscuous, maintaining a monoandrous mating system being less subject to STDs.

We hypothesize that chimpanzees, as a promiscuous species, are likely to be more exposed to STDs (50, 52, 54). After the duplication event that originated *WFDC6*, an episode of rapid evolution may have occurred allowing for the accumulation of amino acid replacements. Later in chimpanzee evolution, the newly originated *WFDC6* appears to have been preserved by strong selective constraints, perhaps representing an adaptive response to a higher load of pathogens. Conversely, in less promiscuous species like orangutan, rhesus and baboon the signals of pseudogenization present on

WFDC6 seem to be associated with more relaxed constraints (higher ω values) and lower pathogen exposure or to the exploitation of different mechanisms of immune defense (50, 51, 53, 55). However, as WFDC6 biological functions and target molecules have not been explored yet, our hypothesis regarding the purifying selective pressures cannot be totally unraveled.

Overall, our data provides support for a clear genetic differentiation of *P. t. verus* from the remaining common chimpanzee's subspecies, and for a single departure of neutrality in the *WFDC* locus in chimpanzees due to strong selective constrains acting on *WFDC6*. We hypothesize that the latter may be due to an adaptive process associated to the expanded antimicrobial spectrum of WFDCs in the male reproductive tract.

Methods

DNA samples and sequence generation

The DNA samples include 68 chimpanzees, including Central African subspecies, *P. t. troglodytes* (15 individuals), Western African subspecies, *P. t. verus* (22 individuals) and the Gulf of Guinea subspecies, *P. t. ellioti* (31 individuals). (Table S.1). We studied the genetic variation in the *WFDC* locus by sequencing the coding regions of 18 *WFDC* and *SEMG* genes (comprising a total of 66 exons) and a number of intervening noncoding regions (spaced every ~10 kb). Additionally, we sequenced 47 pseudogenes located in unrelated, neutrally evolving regions across the chimpanzee genome, used as control regions, as previously used (26, 56) (Table S.2). Primers for amplification and sequencing were designed based on the Human Genome Reference Sequence (March 2006 assembly - v36.1), available at the UCSC Genome Browser (genome.ucsc.edu). All samples were PCR-amplified and analyzed by bidirectional Sanger sequencing. Further details about PCR and DNA sequencing are available from the authors upon request. The sequences were aligned to the Human Genome v26.1 and polymorphic sites and fixed differences were detected with phred-phrap-consed package (57). To ensure sequence quality, we discarded variant sites in the first and last 75 bp of each amplicon segment. We manually curated sites found to have discordant genotypes in different amplicons. The ancestral state of each SNP

was inferred by comparison with the human, orangutan, and macaque genome sequences [(56, 58); genome.ucsc.edu].

Statistical analysis

Summary statistics of population genetic variation were calculated using SLIDER (http://genapps.uchicago.edu/slider/index.html). We assessed statistical significance of summary statistics using an empirical comparison to the control regions, by calculating the upper and lower 2.5 percentiles of each distribution. Specifically, we used the sequenced control regions to perform an empirical comparison of nucleotide diversity (π) and Tajima's *D* values for each *WFDC* gene in each population. As a result, we circumvent the need to specify a history of demographic events shaping the genetic diversity in these three subspecies. For the *WFDC* genes that presented summary statistic values in the tail of the empirical distribution in *P. t. troglodytes* , we ran 10^5coalescent simulations using "ms" (59) and with estimates of mutation rate parameters estimated from the sequenced data with SLIDER. For the population recombination rate we used the PANMAP estimates of chimpanzee recombination. We assumed demographic models that included constant population size and historic events as previously inferred (Jody Hey (2010) and Daniel Wegmann (2010)). For each model, we calculated a null distribution of summary statistics values and calculated the 2.5th and 97.5th percentiles.

We performed HKA test considering all subspecies in DNAsp 5.1 and using a Maximum Likelihood method that incorporates values for multiple neutrally evolving regions (35). MKT was calculated in DNAsp v5.1 using humans as outgroup and assuming two types of sites: putatively neutral sites (Syn) and functional sites (NSyn) (60, 61).

We assessed the subspecies differentiation levels by calculating the F_{ST} statistic and PCA for each SNP (27, 28). We used a locus-by-locus AMOVA, using 20,000 simulations, which was performed by Arlequin using its default values(constant model (62). The EIGENSOFT software package was used for PCA (28). We performed cluster analysis using STRUCTURE version 2.3 software package (30), assuming admixture and correlated allele frequencies. Fifty iterations of the data at each K =1–5 with 500,000 Markov Chain Monte Carlo (MCMC) burn-in steps and 500,000 MCMC

iterations. STRUCTURE output was processed with CLUMPP and plotted with DISTRUCT (63). We used STRUCTURE harvester to determine the best K estimate. Population structure analyses were performed blinded to a priori population labels.

Haplotype phasing for all samples was inferred separately for the *WFDC*-CEN and *WFDC*-TEL sub-loci using PHASE2.1 (64, 65). Haplotypes were independently inputted in Haploview 4.2 (66) to calculate LD statistics, r^2 and *D'*, and to identify LD and haplotype blocks (67). The potential functional effects at the protein level of nonsynonymous SNPs and fixed differences were inferred using PolyPhen-2v (68) and SIFT (69).

Maximum likelihoods estimates of d_N/d_S (ω) were carried out using the *codeml* program from the software package Phylogenetic Analysis by Maximum Likelihood - PAML version 4.2 (39). To run PAML, we first reconstructed a phylogenetic tree (DNAml from Phylogeny Inference Package (PHYLIP - http://evolution.genetics.washington.edu/phylip.html). We used the genomic sequences from human (*Homo sapiens)*, chimpanzee (*Pan troglodytes*), gorilla (*Gorilla gorilla)*, orangutan (*Pongo pygmaeus)*, gibbon (*Nomascus leucogenys),* rhesus monkey (*Macaca mulatta),* baboon (*Papio anubis)*, and marmoset *(Callithrix jacchus)*. They were retrieved from public databases using EPPIN isoform 1 (**Uniprot: O95925**) and WFDC6 isoform 1 (**Uniprot: Q9BQY6) as BLAT templates**. *Pan paniscus* was not included in the analysis as the cDNA sequence is equal to *Pan troglodytes*. The phylogenetic tree diverged from the known primate phylogeny in the position of orangutan and gibbon branches. To test for variable selective pressures among branches, we performed the branch model using either the null model (one ω ratio for the entire tree) or nested models (two-ratios, three-ratios, four-ratios for the tree, defined as depicted in Fig S10) (70, 71). The values of ω > 1 were considered as evidences of positive selection, and the values ω < 1 were considered as an indication of purifying selection. The significance of each model was assessed from twice the variance of likelihoods (-2Δl) using χ^2 statistic.

Acknowledgments

The authors acknowledge Anh-Dao Nguyen for the help inferring the ancestral allele state of each SNP and Riverside Zoo, Sunset Zoo, Lincoln Park Zoo, the Primate

Foundation of Arizona, New Iberia Research Center, and Texas Biomed for sample donation. This work was supported in part by the Intramural Research Program of the National Human Genome Research Institute, by the Portuguese Foundation for Science and Technology (FCT), financed by the European Social Funds and national funds of the Portuguese Ministry of Education and Science (POPH-QREN) fellowship SFRH/BD/45907/2008 to Z.F and grant PTDC/BEX-GMG/0242/2012 to S.S and by the Wellcome Trust Centre for Human Genetics (WT097307) to W.W.K. IPATIMUP is an Associated Laboratory of the Portuguese Ministry of Education and Science and is partially supported by FCT.

References

1. Gonder MK*, et al.* (1997) A new west African chimpanzee subspecies? *Nature* 388(6640):337.
2. Gonder MK*, et al.* (2011) Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. *PNAS* 108(12):12.
3. Patten MA & Unitt P (2002) Diagnosability versus mean differences of sage sparrow subspecies. *The Auk* 119(1):9.
4. Gonder MK, Disotell TR, & Oates JF (2006) New Genetic Evidence on the Evolution of Chimpanzee Populations and Implications for Taxonomy. *International Journal of Primatology* 27(4):1103-1127.
5. Becquet C, Patterson N, Stone AC, Przeworski M, & Reich D (2005) Genetic structure of chimpanzee populations. *PLoS Genetics* preprint(2007):e66.
6. Bowden R*, et al.* (2012) Genomic Tools for Evolution and Conservation in the Chimpanzee: Pan troglodytes ellioti Is a Genetically Distinct Population. *PLoS Genetics* 8(3):e1002504.
7. Keele BF*, et al.* (2006) Chimpanzee Reservoirs of Pandemic and Nonpandemic HIV-1. *Science* 313(5786):3.
8. Van Heuverswyn F*, et al.* (2007) Genetic diversity and phylogeographic clustering of SIVcpzPtt in wild chimpanzees in Cameroon. *Virology* 368(1):155-171.
9. Jones KE*, et al.* (2008) Global trends in emerging infectious diseases. *Nature* 451(7181):990-993.
10. Chimpanzee Sequencing and Analysis Consortium (2005) Initial sequence of the chimpanzee genome and comparison with the human genome. *Nature* 437(7055):69-87.
11. Clauss A, Lilja H, & Lundwall A (2005) The evolution of a genetic locus encoding small serine proteinase inhibitors. *Biochemical and biophysical research communications* 333(2):383-389.
12. Lundwall A (2007) A locus on chromosome 20 encompassing genes that are highly expressed in the epididymis. *Asian journal of andrology* 9(4):540-544.
13. Peter A, Lilja H, Lundwall A, & Malm J (1998) Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase. *Eur. J. Biochem.* 252:5.

14. de Lamirande E (2007) Semenogelin, the main protein of the human semen coagulum, regulates sperm function. *Seminars in thrombosis and hemostasis* 33(1):60-68.
15. Yenugu S*, et al.* (2004) Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif. *Biology of reproduction* 71(5):1484-1490.
16. Bingle CD & Vyakarnam A (2008) Novel innate immune functions of the whey acidic protein family. *Trends in immunology* 29(9):444-453.
17. Clauss A, Persson M, Lilja H, & Lundwall A (2011) Three genes expressing Kunitz domains in the epididymis are related to genes of WFDC-type protease inhibitors and semen coagulum proteins in spite of lacking similarity between their protein products. *BMC Biochem* 12:55.
18. Williams SE, Brown TI, Roghanian A, & Sallenave JM (2006) SLPI and elafin: one glove, many fingers. *Clinical science* 110(1):21-35.
19. Wang Z, Widgren EE, Sivashanmugam P, O'Rand MG, & Richardson RT (2005) Association of eppin with semenogelin on human spermatozoa. *Biology of reproduction* 72(5):1064-1070.
20. Edstrom AM*, et al.* (2008) The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins. *J Immunol* 181(5):3413-3421.
21. Zhao H, Lee WH, Shen JH, Li H, & Zhang Y (2008) Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma. *Peptides* 29(4):505-511.
22. Martellini JA*, et al.* (2009) Cationic polypeptides contribute to the anti-HIV-1 activity of human seminal plasma. *FASEB J* 23(10):3609-3618.
23. Dixson AF & Anderson MJ (2002) Sexual Selection, Seminal Coagulation and Copulatory Plug Formation in Primates. *Folia Primatology* 73:6.
24. Dixson AF & Anderson MJ (2004) Sexual behavior, reproductive physiology and sperm competition in male mammals. *Physiology & behavior* 83(2):361-371.
25. Jensen-Seaman MI & Li WH (2003) Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. *Journal of molecular evolution* 57(3):261-270.
26. Ferreira Z*, et al.* (2013) Reproduction and Immunity-Driven Natural Selection in the Human WFDC Locus. *Mol Biol Evol*.

27. Excoffier L (2002) Human demographic history: refining the recent African origin model. *Current Opinion in Genetics & Development* 12:8.
28. Patterson N, Price AL, & Reich D (2006) Population structure and eigenanalysis. *PLoS Genetics* 2(12):e190.
29. Falush D, Stephens M, & Pritchard JK (2003) Inference of Population Structure Using Multilocus Genotype Data: Linked Loci and Correlated Allele Frequencies. *Genetics* 164:20.
30. Pritchard JK, Stephens M, & Donnelly P (2000) Inference of Population Structure Using Multilocus Genotype Data. *Genetics* 155:14.
31. Tajima F (1989) Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. *Genetics* 123(3):585-595.
32. Fu Y-X & Li W-H (1993) Statistical Tests of Neutrality of Mutations. *Genetics* 133:15.
33. Fay JC & Wu C-I (2005) Detecting hitchhiking from patterns of DNA polymorphism. *Selective sweeps*, ed Nurminsky D (Landes Biosciences, Georgetown, Texas).
34. Zeng K, Fu YX, Shi S, & Wu CI (2006) Statistical tests for detecting positive selection by utilizing high-frequency variants. *Genetics* 174(3):1431-1439.
35. Hudson RR, Kreitman M, & Aguadé M (1987) A Test of Neutral Molecular Evolution Based on Nucleotide Data. *Genetics* 116:6.
36. Wegmann D & Excoffier L (2010) Bayesian inference of the demographic history of chimpanzees. *Molecular Biology and Evolution* 27(6):1425-1435.
37. Hey J (2010) The Divergence of Chimpanzee Species and Subspecies as Revealed in Multipopulation Isolation-with-Migration Analyses. *Molecular Biology and Evolution* 27(4):921-933.
38. Caswell JL, *et al.* (2008) Analysis of chimpanzee history based on genome sequence alignments. *PLoS Genetics* 4(4):e1000057.
39. Yang Z (2007) PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* 24(8):1586-1591.
40. Hurle B, Swanson W, Program NCS, & Green ED (2007) Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. *Genome research* 17(3):276-286.
41. Dorus S, Evans PD, Wyckoff GJ, Choi SS, & Lahn BT (2004) Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. *Nature genetics* 36(12):1326-1329.

42. Kingan SB, Tatar M, & Rand DM (2003) Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. *Journal of molecular evolution* 57(2):159-169.
43. Won YJ & Hey J (2005) Divergence population genetics of chimpanzees. *Molecular Biology and Evolution* 22(2):297-307.
44. Fischer A, Wiebe V, Paabo S, & Przeworski M (2004) Evidence for a complex demographic history of chimpanzees. *Molecular Biology and Evolution* 21(5):799-808.
45. Fischer A, Pollack J, Thalmann O, Nickel B, & Paabo S (2006) Demographic history and genetic differentiation in apes. *Current biology : CB* 16(11):1133-1138.
46. Wilkinson TS, Roghanian A, Simpson AJ, & Sallenave JM (2011) WAP domain proteins as modulators of mucosal immunity. *Biochemical Society transactions* 39(5):1409-1415.
47. Drannik AG, Henrick BM, & Rosenthal KL (2011) War and peace between WAP and HIV: role of SLPI, trappin-2, elafin and ps20 in susceptibility to HIV infection. *Biochemical Society transactions* 39(5):1427-1432.
48. Huhtinen K*, et al.* (2009) Serum HE4 concentration differentiates malignant ovarian tumours from ovarian endometriotic cysts. *British journal of cancer* 100(8):1315-1319.
49. Nunn CL (2002) A comparative study of leukocyte counts and disease risk in primates. *Evolution* 56(1):177–190.
50. Nunn CL, Gittleman JL, & Antonovics J (2000) Promiscuity and the primate immune system. *Science* 290(5494):1168-1170.
51. Nunn CL (2003) Behavioural defences against sexually transmitted diseases in primates☆. *Animal Behaviour* 66(1):37-48.
52. Wlasiuk G & Nachman MW (2010) Promiscuity and the rate of molecular evolution at primate immunity genes. *Evolution* 64(8):2204-2220.
53. Holmes EC (2004) Adaptation and immunity. *PLoS Biol* 2(9):E307.
54. Garamszegi LZ & Nunn CL (2011) Parasite-mediated evolution of the functional part of the MHC in primates. *J Evol Biol* 24(1):184-195.
55. Anderson MJ, Hessel JK, & Dixson AF (2004) Primate mating systems and the evolution of immune response. *J Reprod Immunol* 61(1):31-38.

56. Andrés AM*, et al.* (2010) Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. *PLoS Genetics* 6(10):e1001157.
57. Nickerson DA, Tobe VO, & Taylor SL (1997) PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing. *Nucleic Acids Res* 25(14):2745-2751.
58. Gibbs RA*, et al.* (2007) Evolutionary and biomedical insights from the rhesus macaque genome. *Science* 316(5822):222-234.
59. Hudson RR (2002) Generating samples under a Wright-Fisher neutral model of genetic variation. *Bioinformatics* 18(2):337-338.
60. Rozas J & Rozas R (1995) DnaSP, DNA sequence polymorphism: an interactive program for estimating population genetics parameters from DNA sequence data. *Comput Appl Biosci* 11(6):621-625.
61. Rozas J (2009) DNA sequence polymorphism analysis using DnaSP. *Methods Mol Biol* 537:337-350.
62. Excoffier L, Laval G, & Schneider S (2005) Arlequin (version 3.0): An integrated software package for population genetics data analysis. *Evolutionary Bioinformatics Online* 1:3.
63. Conrad DF*, et al.* (2006) A worldwide survey of haplotype variation and linkage disequilibrium in the human genome. *Nature genetics* 38:1251-1260.
64. Stephens M, Smith NJ, & Donnelly P (2001) A new statistical method for haplotype reconstruction from population data. *Am J Hum Genet* 68(4):978-989.
65. Stephens M & Donnelly P (2003) A comparison of bayesian methods for haplotype reconstruction from population genotype data. *Am J Hum Genet* 73(5):1162-1169.
66. Barrett JC (2009) Haploview: Visualization and analysis of SNP genotype data. *Cold Spring Harb Protoc* 2009(10):pdb ip71.
67. Gabriel SB*, et al.* (2002) The structure of haplotype blocks in the human genome. *Science* 296(5576):2225-2229.
68. Adzhubei IA*, et al.* (2010) A method and server for predicting damaging missense mutations. *Nat Methods* 7(4):248-249.
69. Kumar P, Henikoff S, & Ng PC (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nature Protocols* 4(7):1073-1082.

70. Yang Z (1997) PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* 13(5):555-556.

71. Bielawski JP & Yang Z (2003) Maximum likelihood methods for detecting adaptive evolution after gene duplication. *Journal of Structural and Functional Genomics* 3:11.

Table 1: Summary Statistics for all the *WFDC* genes in *Pan troglodytes*.

Gene	[a]L	[b]S	[c]π	[d]Θ_w	[e]*D*	[f]*D**	[g]*H*	[h]*P* (HKA)
WFDC5	5536	45	7.78	8.25	-0.174	-0.182	1.41	0.675
WFDC12	1323	24	2.48	4.39	-1.25	-1.08	0.723	0.664
PI3	3377	32	4.32	5.85	-0.776	-0.378	6.52	0.445
SEMG1	3305	45	4.00	8.20	-1.57	0.336	-1.04	0.146
SEMG2	4324	47	3.58	8.57	-1.78	-0.352	5.07	0.816
SLPI	4709	46	6.68	8.38	-0.621	0.639	7.66	0.247
WFDC2	3984	34	5.35	6.2	-0.406	-0.248	0.656	0.104
SPINT3	3727	44	5.30	8.04	-1.04	-1.59	36.9	0.008
WFDC6	2807	31	1.63	5.67	-2.10	-2.90	-1.52	0.013
WFDC7	3233	38	3.16	6.96	-1.65	-2.09	16.9	0.018
WFDC8	7179	56	4.82	10.2	-1.64	-2.18	6.13	0.115
WFDC9/10A	6863	67	7.24	12.2	-1.28	-1.86	19.4	0.386
WFDC11	5037	71	10.1	13.0	-0.708	-0.529	25.8	0.014
WFDC10B/13	7365	59	6.85	10.8	-1.13	-1.54	-4.13	0.105
SPINT4	3527	30	2.21	5.50	-1.76	-1.95	16.1	0.173
WFDC3	7572	93	10.6	17.0	-1.19	-1.02	17.8	0.071

[a] L = Legth sequenced (bp).

[b] S = number of segregating sites.

[c] Nucleotide diversity per base pair (x 10^{-4})

[d] Watterson's estimator of Θ ($4N_e\mu$) (Watterson 1975) per base pair (x 10^{-4}).

[e] *D* = Tajima's *D* statistic (Tajima 1989).

[f] *D** = Fu and Li's *D** test (Fu and Li 1993)

[g] Fay and Wu *H* test (Fay and Wu 2002; Zeng et al. 2006)

[h] HKA test *P*-value (Hudson, Kreitman and Aguadé 1987)

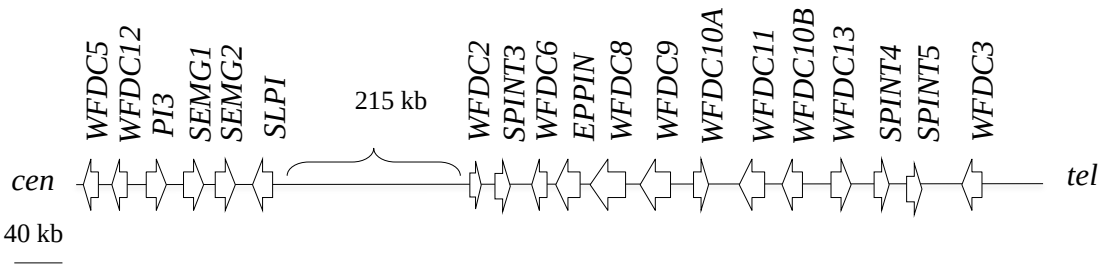


Figure 1: Schematic representation of the 20q13 *WFDC* gene locus, showing the relative positions of the *WFDC* genes. As depicted, the *WFDC* locus spans 700 kb and its genes are organized into two subloci (centromeric and telomeric; *WFDC*-CEN and *WFDC*-TEL, respectively), separated by 215 kb of unrelated sequence.

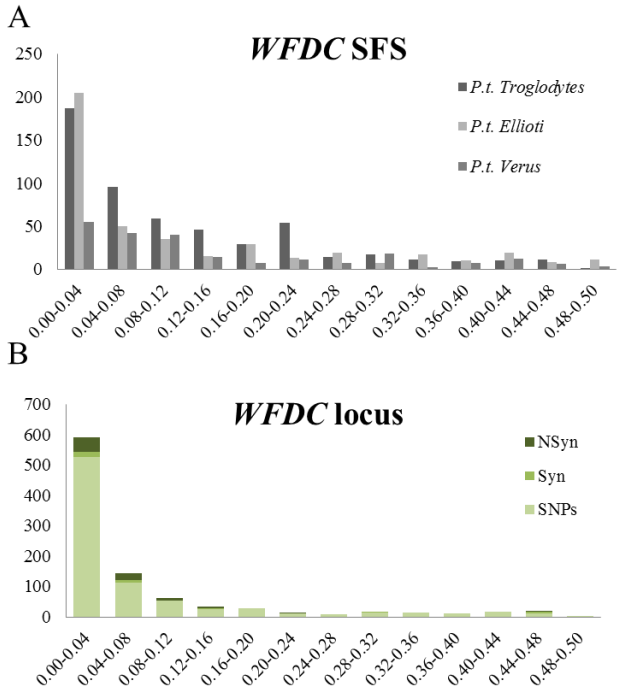


Figure 2: Folded site frequency spectrum (folded SFS) for the species resequenced. The *x*-axis depicts the frequency of the allele frequency bin in the generated data set, whereas the y axis represents the number of alleles found within each frequency bin. Syn, synonymous changes; NSyn, nonsynonymous changes. (A) folded SFS in *WFDC* locus; (B) folded SFS of *WFDC* locus highlighting coding mutations.

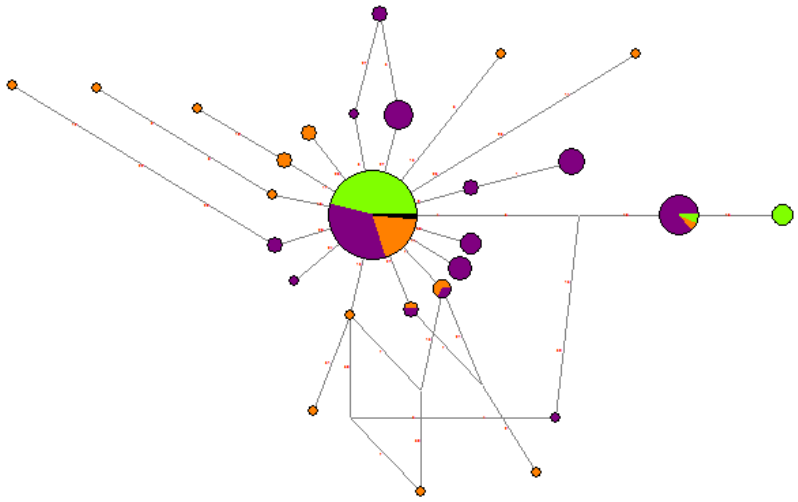


Figure 3: Inferred network haplotypes at the *WFDC6*. Each circle represents a unique haplotype, and its area is proportional to its frequency. Within each circle, *P. t. verus*, *P. t. ellioti*, and *P. t. troglodytes* are labeled in green, purple and orange, respectively. The mutations that differentiate each haplotype are shown along each branch. The inferred network haplotype of *WFDC6* shows a star-like structure characteristic of a population expansion, combined with purifying selection.

3.3. **Biological assessment of candidate genes**

3. 3. 1. Characterization of the Human *WFDC8*: Evolutionary history and differential allele expression

In preparation

Characterization of the Human *WFDC8*: Evolutionary history and differential allele expression

Zélia Ferreira[1, 2], Carla Bartosch[3], Fátima Carneiro[1, 3], Susana Seixas[1]

1 - Institute of Molecular Pathology and Immunology of the University of Porto (IPATIMUP), Porto 4200-465, Portugal;

2 - Department of Zoology and Anthropology, Faculty of Sciences, University of Porto, Porto 4099-002, Portugal;

3 – Service of Pathology, São João Hospital, Porto 4200-319, Portugal

Corresponding author: Susana Seixas,

IPATIMUP

Rua Dr. Roberto Frias s/n, 4200-465 Porto, Portugal

keywords: WFDC8, serine protease inhibitor, ASE

Running head: Differencial allelic expression of *WFDC8*.

Abstract

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) locus located on human chromosome 20q13, spans 17 *WFDC* genes with WAP and/or Kunitz domains. The WFDCs are small-secreted molecules known for their roles in innate immunity, reproduction, and regulation of endogenous proteases. *WFDC8* was pinpointed as target of balancing selection in the Europeans. The putative candidate variant -44A (rs7273669A) is located near the 5'UTR, and was predicted to affect gene expression by abolishing the binding site of two transcription factors. We explored the activity of WFDC8 in humans, by studying its conservation across 12 mammalian species, and by analyzing patterns of expression linked to -44A allele. We show that *WFDC8* has been conserved throughout mammalian evolution and we provide *in vivo* and *in vitro* evidence for an effect of -44(G/A) on gene expression. Furthermore, we propose a differential allele expression dependent of the cellular and/or tissue background. The divergent expression effects of -44(G/A) in different tissues strengthens the balancing selection hypothesis, highlighting the balanced effect of the polymorphisms in *WFDC8* expression in human tissues.

Introduction

The whey acidic protein (WAP) four-disulfide core domain (*WFDC*) locus, located on human chromosome 20q13, encodes small-secreted peptides, with roles in innate immunity, reproduction, and regulation of endogenous proteases [1-6].

The majority of WFDC molecules contain one or more WAP domains and a few of these proteins include other domains such as trappin (e. g. PI3); or Kunitz protease inhibitor domains (e. g. EPPIN and WFDC8) [7, 2]. Peptidase inhibitor 3 (*PI3,* also known as *elafin*) and secretory leukocyte peptidase inhibitor (*SLPI*) are the most studied members of the *WFDC* locus and encode pleiotropic proteins involved in recognizing microbial infections at mucosal surfaces and are also thought to be involved in sperm maturation [8-11]. *SEMG1* and *SEMG2* are nearby genes that share conserved flanking regions with *WFDC*s, and encode the main protein components of the semen coagulum. SEMGs have combined roles in male reproduction and pathogen response; when crosslinked SEMGs entrap spermatozoa in the coagulum and when cleaved into small peptides SEMGs present antimicrobial activity [6,12-14]. Another example of a thoroughly studied WFDC member is EPPIN, which binds to SEMG1 and protects it from premature cleavage by its natural protease – Prostate Specific Antigen (PSA, or kallikrein-3). EPPIN has remarkable implications in reproductive biology since it participates in spermatozoa maturation and in their protection from microbial attack [3,12,13,15-17]. To our knowledge, the biological properties of the other *WFDC* genes remain poorly characterized as it is the case of *WFDC8*. This gene encodes a protein with three WAP domains and one Kunitz domain, which is expected to combine antimicrobial and antiprotease properties [5]. Its rat homologue *Wfdc8,* was shown to present a gene expression dependent of androgens and inducted by lipopolysaccharides (LPS) and its protein to display a potent antibacterial activity [18].

In humans WFDC8 was identified in the semen [19,20], the loss of its gene expression was associated with a fertility impairment after vas deferent renamastosis and in Hutterite pedigrees *WFDC8* appears to be associated with transmission distortion [21]. Moreover, *WFDC8* was previously shown to display a signature of balancing selection in Europeans associated to a -44(G/A) polymorphism (rs7273669), located in the 5’ region of *WFDC8* gene. This polymorphism was predicted to alter gene expression by abolishing the binding site of two transcription factors [22,23]. Balancing selection is an evolutionary process in which genetic variation is preserved

throughout time and it is generally associated with an advantage of heterozygotes. Such examples include carriers of sickle cell anemia, which are resistant to *Plasmodium falciparum* infection (malaria) and the large inter individual variation of MHC locus, which protects humans against a large spectrum of pathogens [24-31]. Importantly, signals of balancing selection appear to be more common in genes that are involved in host-pathogen interactions and immune response [32-37].

Here, we describe the patterns of *WFDC8* conservation and the frequency distribution of -44(G/A) polymorphism in different populations of European origin and we evaluate the impact of -44(G/A) polymorphism on *WFDC8* expression by means of allele-specific amplification and luciferase assays. We demonstrate that *WFDC8* has been conserved throughout mammalian evolution and we provide evidence for an effect of -44(G/A) on gene expression that depends on the cellular background.

Methods

Phylogenetic analysis

WFDC8 cDNA sequences were extracted from public databases (Supplementary Table S1) and their phylogeny was built using the maximum likelihood method implemented in the DNAml program from the Phylogeny Inference Package (PHYLIP – http://evolution.genetics.washington.edu/phylip.html). Maximum likelihood estimates of d_N/d_S, were carried out using the *codeml* program from the software package Phylogenetic Analysis by Maximum Likelihood - PAML version 4.5. The following likelihood rate tests (LRTs) were performed: 1) the *branch* model which compares a single d_N/d_S value obtained for all lineages (M0) with a model assuming different d_N/d_S values for each lineage (free-ratio); 2) the *site* models, which allow the d_N/d_S value to vary among sites of the protein and compare models of neutrality with positive selection (M1-M2 and M7-M8). Values of $d_N/d_S > 1$ were considered as evidence of positive selection and the values of $d_N/d_S < 1$ were regarded as an indicative of purifying selection. The significance of each nested model was obtained from twice the variation of likelihoods ($2\Delta l$) using a χ^2 statistic. The Bayes empirical Bayes (BEB) was used to calculate posterior probabilities of site classes, in order to identify sites under positive selection for the significant LRTs [38].

Tissue Expression

To investigate the pattern of distribution of *WFDC8* we analyzed 19 cDNA samples from different healthy organs. Except for the first-strand cDNA from leukocytes (Clontech), the tissue cDNA samples were synthesized by reverse transcriptase methods using as templates the RNA from the First Choice Human Total RNA Survey Panel (Ambion). Reverse transcription was performed using the Superscript III RT PCR system (Life Technologies) according to the manufacturer's protocol. The primers used for amplification of the cDNA of *WFDC8* were 5'- GGA GTG GAC TTC TGC AAT GCT GA-3' and 3'- TCA TGA GGC ACA GCG CTG GCT-5'. A segment of the house-keeping gene *GAPDH* was employed as internal control using the forward primer 5'-TCA AGG CTG AGA ACG GGA AG -3' and the reverse primers 5'-AGA GGG GGC AGA GAT GAT GA-3' and 3'-CAG TAG AGG CAG GGA TGA TG-5'. For amplification,

we used the following cycling conditions: 95 ºC for 15 min and 35 cycles of 94 ºC for 30 s, 62 ºC for 30 s and 72 ºC for 1 min and extension time for 30 min at 60 ºC. Amplification fragments were separated by 9% polyacrylamide gel electrophoresis.

Allele-specific expression in Ovarian Tissues

Genomic DNA (gDNA) and total mRNA was extracted from 12 frozen ovarian samples donated by the Tumor and Tissue Bank from Saint John Hospital (Porto, Portugal). For the extraction of gDNA we used QIAamp DNA Mini and Blood Mini kit (Qiagen) following the manufacturer's instructions. For the purification of total mRNA we used first Tripure Isolation reagent (Roche Applied Science). Then, to remove possible gDNA contamination, mRNA was treated with DNase I (Ambion) and precipitated with acidic phenol-chloroform (Ambion). Double stranded cDNA was synthesized from total RNA using the reverse transcriptase protocol referred above.

To evaluate the impact of -44(G/A) polymorphism in gene expression we genotyped the marker rs2250860 located in the *WFDC8* coding region. This SNP was selected due to its strong association (D'= 1 and r^2=1) with -44(G/A) polymorphism in a sample of individuals with ancestry in northern or western Europe (CEU sample from HapMap project) and it was used as a proxy for allele-specific expression. To certify the association of rs2250860 and -44(G/A) in our sample, we Sanger-sequenced both SNPs using the gDNA extracted from the 12 ovarian tissues. TaqMan Gene Expression Assay design for rs2250860 (Applied Biosystems) and TaqMan PCR technology was used in experiments of allele-specific expression. The analysis was carried out in the cDNA samples from heterozygotes individuals and a gDNA sample from a heterozygous individual was used as reference for 50:50 ratio. Reactions were prepared as recommended for the genotyping assay, using Taqman master mix (Applied Biosystems), the two primers,and FAM- and VIC labeled probes, each designed to specifically anneal to one of the alleles of rs2250860. The genotyping was

run on an ABI PRISM 7500 RT-PCR system. The experiments were repeated at least twice. The same protocol was used to quantify the differential levels of allelic expression for each sample, with the addition of a series of dilutions from a gDNA sample of a heterozygous were used to generate a standard curve. A threshold of 35 Ct was used to consider allele expression. Relative expression quantitation was extrapolated from linear regression for the single sample presenting both alleles expression above the 35 Ct. This experiment was run in triplicate.

Plasmid construction

A segment of the 5' *WFDC8* region (positions -351 to 115 relative to the translation start site) was amplified by PCR from individuals with known genotypes (AA and GG). The PCR was performed using the primers 5'- ACA TTT AGC AAG GTA GGC GT-3' and 5'-GAC CAC TCT AAC TTC TAT GT-3'. The PCR product was first cloned into a pCR 4-TOPO-TA cloning vector (Life Technologies), digested with *SacII* and *XhoI* and ligated 5' of the Luciferase sequence of the pGL3-Promoter vector. All plasmids were sequenced before transfection.

Cell culture

Human Cervix Adenocarcinoma (HeLa) and Colorectal Cancer cell lines (Caco-2,) were cultured in Dulbecco's Modified Eagle Medium (DMEM; Life Technologies) supplemented with 10% fetal bovine serum (FBS; Life Technologies) and penicillin/streptomycin antibiotics.

Transient transfection

Cells (~10^5) were seeded into a 24-well plate 24 hours before transfection. Plasmid constructs were co-transfected with phRL-TK (Promega) at a 2:1 ratio using lipofectamine agent (Life Technologies). Twenty-four hours after transfection, cells were harvested and luciferase activity was measured by Dual-Luciferase Reporter Assay System (Promega) according to the manufacturer's protocol. The result was expressed as the ratio between firefly and *Renilla luciferase*. Fifteen replicates were performed for each experiment. Independent t-tests were performed in SPSS 15.0 (SPSS, Inc., Chicago, IL) and the null hypothesis was rejected when $P < 0.05$.

Results

Evolutionary study of WFDC8

To investigate the evolution of *WFDC8* and its levels of conservation, we performed a series of phylogenetic analyses using the coding sequences of 12 mammalian species: *Homo sapiens*, *Pan troglodytes*, *Pan paniscus*, *Gorilla gorilla*, *Pongo abelii*, *Nomascus leucogenys*, *Papio anubis*, *Macaca mulatta, Saimiri boliviensis, Mus musculus*, *Rattus norvegicus*, and *Canis lupus familiaris*. The organization of WFDC8 in 3 WAP and 1 Kunitz domains is conserved in all species except for *P. abelii,* in which the terminal WAP domain was lost (Figure 1A). The 30 cysteines present in the *H. sapiens* sequence are conserved in the remaining species pointing to the maintenance of the 15 disulfide bridges in the WFDC8 molecule throughout mammalian evolution (Figure 1A).

To explore the nature of the selective pressures acting on *WFDC8* we calculated d_N/d_S ratios (d_N and d_S correspond to non-synonymous and synonymous substitution rates, respectively) assuming opposite evolutionary scenarios. The low d_N/d_S estimate (0.4709) obtained for the 12 mammals phylogeny suggests an overall conserved evolution of *WFDC8*. We also tested the hypothesis of some codon positions were evolving under positive selection in *WFDC8*. Although most sites are constrained, a few amino acid positions (2.6 %) are likely to be under positive selection in WFDC8 (34R, 45P, 73T, 139E, 156K, 171D and 252K) including the one located in protease target region of the Kunitz domain (Table 1; Figure 1A).

Frequency distribution of -44G/A polymorphism

To determine the frequency distribution of the -44G/A polymorphism in a larger set of individuals with European ancestry we inquired the 1000 genomes Project database (http://www.1000genomes.org). In the total sample comprising 379 Europeans individuals the frequency of heterozygotes (48.8 %) is slightly higher than expected but not enough to cause a departure from the Hardy-Weinberg equilibrium (HWE). The split of samples by populations (CEU, TSI, GRB, FIN and IBS) and sex uncovered some level of differentiation in heterozygotes frequencies where only the

British females (GBR) show departure from HWE (*P*-value = 0.01) (Supplementary Figure 1).

Tissue expression of WFDC8

To investigate *WFDC8* expression a panel of 19 human tissues (cDNA) was screened by PCR. Our results show that the expression of *WFDC8* is higher in testis, residual in bladder, brain and ovaries and absent in the remaining tissues (Figure 2).

Allele-specific Expression of WFDC8

To test the differential expression of –44G/A polymorphism we sequenced two first strand cDNA libraries for *WFDC8*, a first library from ovaries (normalized pool of 15 individual samples with European ancestry; Clontech) and another one from testis (normalized pool of 45 samples with European ancestry; Clontech). Given the observed frequencies of -44A and -44G alleles (~50%) the analysis of the surrogate marker rs2205806 suggested an expression bias towards -44A allele in ovaries (allele 1 or G) but not in testis (Figure 3A). Within the 12 ovarian samples, we identified nine heterozygous for –44G/A polymorphism, all of which were heterozygous for rs2205806 (Supplementary Table S2). The cDNA samples from the nine heterozygotes were selected for the TaqMan rs2205806 assay to confirm the biased expression of -44A allele. Despite the low expression of *WFDC8* in ovaries, which prevented the amplification of rs2205806 marker in three samples, the remaining samples showed an imbalance towards -44A allele (Figure 3B). The quantification of relative allele expression in the sample with better signal pointed to a three times higher expression of -44A allele.

To further characterize the functional effect of -44G/A polymorphism, we constructed two report vectors -44G and -44A and transfected two cell lines: HeLa and Caco-2. The two cell lines yield contrasting results, while in HeLa -44A allele leads to an expression increase, in Caco-2 cells the same allele reduces *WFDC8* expression (Figure 4).

Discussion

Many genes with roles in immunity have evolved under balancing selection to ensure their function and variability to respond to pathogen infections. WFDC proteins contain different domains, which are thought to confer antimicrobial properties to protect host from pathogen infections and spermatozoa from bacterial attack [2]. Most of the *WFDC*s with known biological functions have combined activities in immunity and reproduction. For example, *EPPIN* encodes for a protein with an active role in protecting spermatozoa from bacterial infections while helping to maturate spermatozoa for fertilization. For the particular case of *WFDC8,* no specific biological function has been attributed, nor has a protease inhibitor activity been cataloged. However, accordingly to the recent findings made for its rat homologue, WFDC8 is likely to have an important role in host-pathogen interaction and male fertility [18].

Comparative and phylogenetic analysis of *WFDC8* in 12 mammalian species showed that all cystein residues and disulfide bridges were preserved during evolution. Overall, *WFDC8* has been conserved except for a few amino acids under positive selection. This is the case of E139 residue, located in the reactive center of the Kunitz domain, which suggests a fast evolution of WFDC8 to counteract the activity of different proteases. Conversely, the maintenance of the 15 disulfide bridges in WFDC8 indicates a possible conservation of antimicrobial function in most mammalians.

The signature of balancing selection reported for Europeans in *WFDC8* is characterized by the occurrence of two intermediate frequency haplotypes (approximately 50%), which are defined by opposite alleles of the -44G/A polymorphism (rs7273669). The haplotype carrying -44A allele was found to extend around 200 kb in the *WFDC* locus as if it had recently arisen in human history. Notwithstanding, the frequency distribution of the -44G/A polymorphism in the larger set of individuals with European ancestry is close to 50% and did not differ significantly among populations. Even with such larger sample, we could not detect an excess of heterozygotes as theoretically predicted for an event of balancing selection. This lack of significance may indicate that the sample size is not enough to detect a bias in heterozygous frequency or that selective effect is not associated with an heterozygous advantage. However, if there is indeed a transmission distortion associated to one of the *WFDC8* haplotypes, it would be expected to find an excess of heterozygous due to

bias in the segregation of one allele. Nonetheless, the sample size in Hutterite pedigrees is not large enough to confidently support this deviation and thus the result could be due to chance fluctuations in transmission rate [21].

The analysis of tissue patterns confirmed that *WFDC8* is manly expressed in testis, though it can be also found at lower levels in other tissues such as ovaries, bladder, and brain. Importantly, for the ovaries we demonstrate that *WFDC8* expression is affected by genotype, in which the -44A allele is up to 3 times more expressed than -44G allele. Although we could not analyze other tissues, this finding may not hold true for other tissues since no hint of allelic imbalance was detected in the testis cDNA library resulting from the pooling of 45 samples with European ancestry. Accordingly, the *in vitro* experiments performed for the opposite allele configurations -44A and -44G yielded contrasting results depending of the cellular model used. In HeLa cells, the -44A allele led to increased expression as observed for the in vivo model of ovaries. Conversely, in Caco-2 cells the -44A allele led to a reduced expression. Altogether, these findings suggest that the same variant -44A might have contrasting effects on *WFDC8* expression depending on cellular or tissue background.

In an evolutionary perspective, *WFDC8* might be under balancing selection due to a fine tuned regulation of gene expression in which being a heterozygous may be advantageous or, on the other hand, the slight increase expression in ovaries can have a beneficial effect in female fertility by conferring an higher antimicrobial potential to female reproductive system without compromising the activity of proteases with a role in fertilization.

In conclusion, WFDC8 is highly conserved in mammalian lineages highlighting the possibility of an important role in host-pathogen interaction and fertility. Further assessment of the biological role of *WFDC8* cataloging its protease inhibitory and antibacterial activity might elucidate on the heterozygote advantage in European populations.

References

[1] A. Lundwall, A. Clauss, Identification of a novel protease inhibitor gene that is highly expressed in the prostate, Biochemical and biophysical research communications 290 (2002) 452-456.

[2] A. Clauss, H. Lilja, A. Lundwall, The evolution of a genetic locus encoding small serine proteinase inhibitors, Biochemical and biophysical research communications 333 (2005) 383-389.

[3] S. Yenugu, R.T. Richardson, P. Sivashanmugam, Z. Wang, G. O'Rand M, F.S. French, S.H. Hall, Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif, Biology of reproduction 71 (2004) 1484-1490.

[4] D. Bouchard, D. Morisset, Y. Bourbonnais, G.M. Tremblay, Proteins with whey-acidic-protein motifs and cancer, The Lancet Oncology 7 (2006) 167-174.

[5] C.D. Bingle, A. Vyakarnam, Novel innate immune functions of the whey acidic protein family, Trends in immunology 29 (2008) 444-453.

[6] A. Lundwall, A. Clauss, Genes encoding WFDC- and Kunitz-type protease inhibitor domains: are they related?, Biochemical Society transactions 39 (2011) 1398-1402.

[7] L. Bingle, S.S. Cross, A.S. High, W.A. Wallace, D. Rassl, G. Yuan, I. Hellstrom, M.A. Campos, C.D. Bingle, WFDC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung, Respiratory research 7 (2006) 61.

[8] S.E. Williams, T.I. Brown, A. Roghanian, J.M. Sallenave, SLPI and elafin: one glove, many fingers, Clinical science 110 (2006) 21-35.

[9] S. Weldon, N. McGarry, C.C. Taggar, N.G. McElvaney, The role of secretory leucoprotease inhibitor in the resolution of inflammatory responses, Biochemical Society Transactions 35 (2007) 3.

[10] S. Weldon, C.C. Taggart, Innate host defense functions of secretory leucoprotease inhibitor, Experimental lung research 33 (2007) 485-491.

[11] P.J. McKiernan, N.G. McElvaney, C.M. Greene, SLPI and inflammatory lung disease in females, Biochemical Society transactions 39 (2011) 1421-1426.

[12] A.M. Edstrom, J. Malm, B. Frohm, J.A. Martellini, A. Giwercman, M. Morgelin, A.M. Cole, O.E. Sorensen, The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins, J Immunol 181 (2008) 3413-3421.

[13] F. Bourgeon, B. Evrard, M. Brillard-Bourdet, D. Colleu, B. Jegou, C. Pineau, Involvement of semenogelin-derived peptides in the antibacterial activity of human seminal plasma, Biology of reproduction 70 (2004) 768-774.

[14] J.A. Martellini, A.L. Cole, N. Venkataraman, G.A. Quinn, P. Svoboda, B.K. Gangrade, J. Pohl, O.E. Sorensen, A.M. Cole, Cationic polypeptides contribute to the anti-HIV-1 activity of human seminal plasma, FASEB J 23 (2009) 3609-3618.

[15] M. Robert, C. Gagnon, Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein, Cell. Mol. Life Sci. 55 (1999) 16.

[16] Z. Wang, E.E. Widgren, P. Sivashanmugam, M.G. O'Rand, R.T. Richardson, Association of eppin with semenogelin on human spermatozoa, Biology of reproduction 72 (2005) 1064-1070.

[17] H. Zhao, W.H. Lee, J.H. Shen, H. Li, Y. Zhang, Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma, Peptides 29 (2008) 505-511.

[18] A. Rajesh, G. Madhubabu, S. Yenugu, Identification and characterization of Wfdc gene expression in the male reproductive tract of the rat, Molecular reproduction and development 78 (2011) 633-641.

[19] I. Batruch, I. Lecker, D. Kagedan, C.R. Smith, B.J. Mullen, E. Grober, K.C. Lo, E.P. Diamandis, K.A. Jarvi, Proteomic analysis of seminal plasma from normal volunteers and post-vasectomy patients identifies over 2000 proteins and candidate biomarkers of the urogenital system, J Proteome Res 10 (2011) 941-953.

[20] D. Kagedan, I. Lecker, I. Batruch, C. Smith, I. Kaploun, K. Lo, E. Grober, E.P. Diamandis, K.A. Jarvi, Characterization of the seminal plasma proteome in men with prostatitis by mass spectrometry, Clin Proteomics 9 (2012) 2.

[21] W.K. Meyer, B. Arbeithuber, C. Ober, T. Ebner, I. Tiemann-Boege, R.R. Hudson, M. Przeworski, Evaluating the evidence for transmission distortion in human pedigrees, Genetics 191 (2012) 215-232.

[22] Z. Ferreira, B. Hurle, J. Rocha, S. Seixas, Differing evolutionary histories of WFDC8 (short-term balancing) in Europeans and SPINT4 (incomplete selective sweep) in Africans, Molecular Biology and Evolution 28 (2011) 2811-2822.

[23] Z. Ferreira, S. Seixas, A.M. Andres, W.W. Kretzschmar, J.C. Mullikin, P.F. Cherukuri, P. Cruz, W.J. Swanson, N.C.S. Program, A.G. Clark, E.D. Green, B. Hurle, Reproduction and Immunity-Driven Natural Selection in the Human WFDC Locus, Mol Biol Evol (2013).

[24] Y. Koda, H. Tachida, M. Soejima, O. Takenaka, H. Kimura, Ancient origin of the null allele se(428) of the human ABO-secretor locus (FUT2), J Mol Evol 50 (2000) 243-248.

[25] B.C. Verrelli, S.A. Tishkoff, A.C. Stone, J.W. Touchman, Contrasting histories of G6PD molecular evolution and malarial resistance in humans and chimpanzees, Molecular Biology and Evolution 23 (2006) 1592-1601.

[26] M. Fumagalli, M. Sironi, U. Pozzoli, A. Ferrer-Admetlla, L. Pattini, R. Nielsen, Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution, PLoS Genetics 7 (2011) e1002355.

[27] G. Wlasiuk, M.W. Nachman, Adaptation and constraint at Toll-like receptors in primates, Molecular Biology and Evolution 27 (2010) 2172-2186.

[28] L.B. Barreiro, L. Quintana-Murci, From evolutionary genetics to human immunology: how selection shapes host defence genes, Nature reviews. Genetics 11 (2010) 17-30.

[29] L.Z. Garamszegi, C.L. Nunn, Parasite-mediated evolution of the functional part of the MHC in primates, Journal of evolutionary biology 24 (2011) 184-195.

[30] H. Areal, J. Abrantes, P.J. Esteves, Signatures of positive selection in Toll-like receptor (TLR) genes in mammals, BMC evolutionary biology 11 (2011) 368.

[31] A.K. Azad, W. Sadee, L.S. Schlesinger, Innate immune gene polymorphisms in tuberculosis, Infect Immun 80 (2012) 3343-3359.

[32] B.C. Verrelli, J.H. McDonald, G. Argyropoulos, G. Destro-Bisol, A. Froment, A. Drousiotou, G. Lefranc, A.N. Helal, J. Loiselet, S.A. Tishkoff, Evidence for balancing selection from nucleotide sequence analyses of human G6PD, Am J Hum Genet 71 (2002) 1112-1128.

[33] F. Prugnolle, A. Manica, M. Charpentier, J.F. Guegan, V. Guernier, F. Balloux, Pathogen-driven selection and worldwide HLA class I diversity, Curr Biol 15 (2005) 1022-1027.

[34] L. Abi-Rached, A.K. Moesta, R. Rajalingam, L.A. Guethlein, P. Parham, Human-specific evolution and adaptation led to major qualitative differences in the variable receptors of human and chimpanzee natural killer cells, PLoS Genetics 6 (2010) e1001192.

[35] A.M. Andres, M.Y. Dennis, W.W. Kretzschmar, J.L. Cannons, S.Q. Lee-Lin, B. Hurle, P.L. Schwartzberg, S.H. Williamson, C.D. Bustamante, R. Nielsen, A.G. Clark, E.D. Green, Balancing selection maintains a form of ERAP2 that undergoes

nonsense-mediated decay and affects antigen presentation, PLoS Genetics 6 (2010) e1001157.

[36] W. Ferguson, S. Dvora, R.W. Fikes, A.C. Stone, S. Boissinot, Long-term balancing selection at the antiviral gene OAS1 in central African chimpanzees, Molecular Biology and Evolution 29 (2012) 1093-1103.

[37] E.M. Leffler, Z. Gao, S. Pfeifer, L. Ségurel, A. Auton, O. Venn, R. Bowden, R. Bontrop, J.D. Wall, G. Sella, P. Donnelly, G. McVean, M. Przeworski, Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees, Science (2013).

[38] Z. Yang, PAML 4: Phylogenetic Analysis by Maximum Likelihood, Mol Biol Evol 24 (2007) 1586-1591.

Table 1: Parameter Estimates and Likelihood Scores under different models.

	-2ΔI Model Comparisons			
	Free Ratio vs M0	M0 vs. M3	M1 vs. M2	M7 vs. M8
WFDC8	68.4327	254.5853	16.34434	23.76999
	Parameter Estimates under M8			**Positively Selected Sites**
	ω (p1) = 0.2217	β (p0) = 0.7783	34R; 45P; 73T; 139E; 156K; 171D; 252E	

Figure 1: Alignment of WFDC8 in 12 mammalian species. Cysteines are marked in light green. Sites positively selected sites with posterior probabilities >0.95 are indicated in red. WAP domains are marked orange, Kunitz domain is marked purple and WFDC8 reactive center is highlighted with the red box.

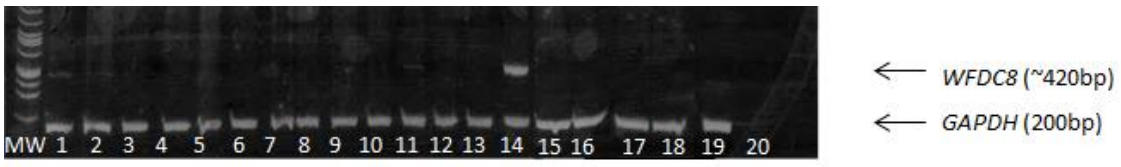


Figure 2: *WFDC8* expression patterns in tissue panel. Lane: MW – Molecular Weigh 1- Bladder; 2- Brain; 3- Colon; 4- Heart; 5- Liver; 6- Adipose Tissue; 7- Cervix; 8- Esophagus; 9- Kidney; 10- Lung; 11- Ovary; 12- Prostate; 13- Small Intestine; 14- Testis; 15- Placenta; 16- Skeletal Muscle; 17- Spleen; 18- Thymus; 19- Leukocyte 20- Negative Control.

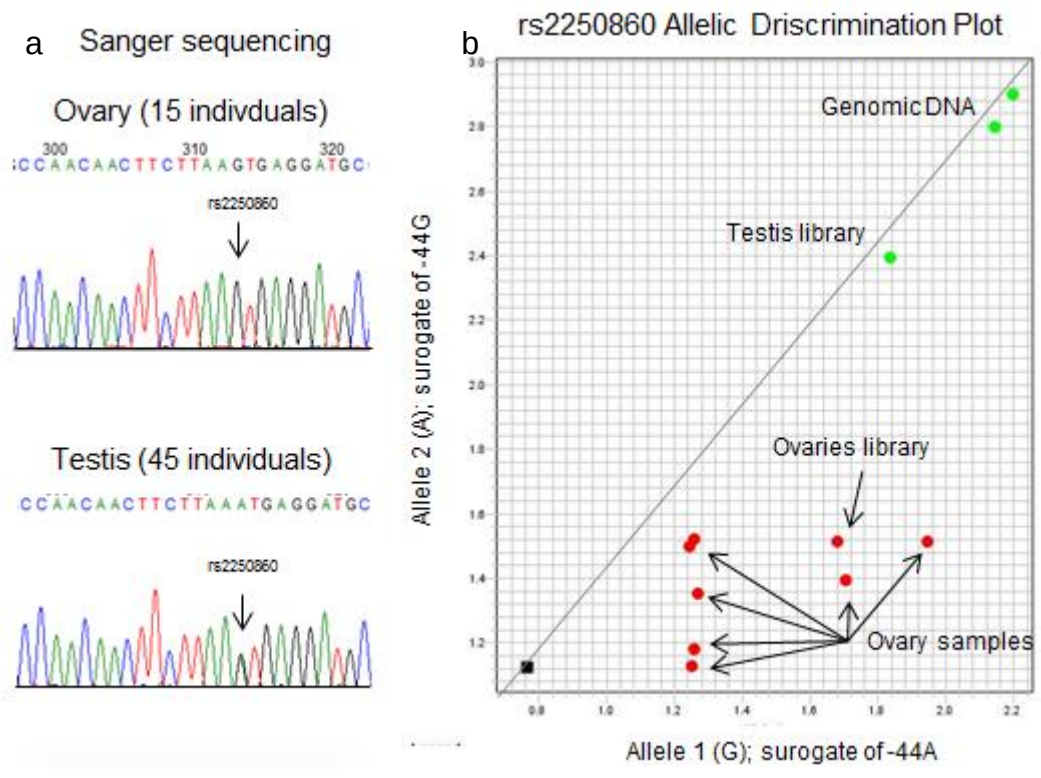


Figure 3: *WFDC8* allelic expression in testis and ovaries (a) Sanger sequencing of ovaries (above) and testis (below) cDNA library; (b) Pairwise correlation analysis of the mean log2 allelic expression ratios for ovary samples, cDNA libraries, and Genomic DNA. Green dots represent 50:50 ratios; red dots represent higer expression of allele 1 (-44A) and the black dot the negative control.

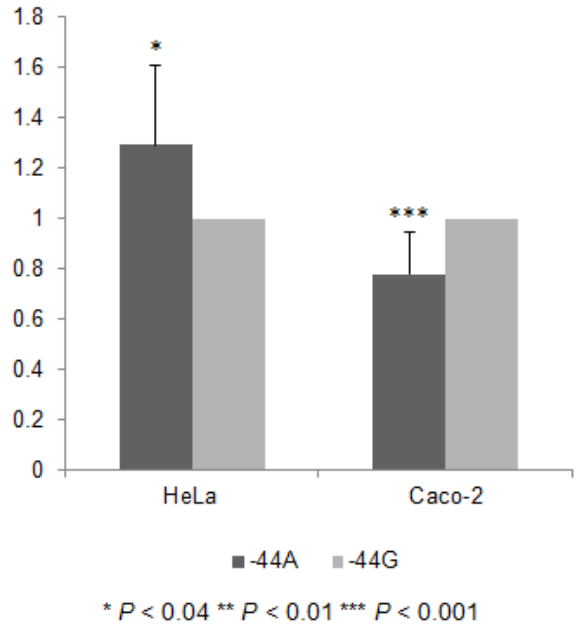


Figure 4: Luciferase activity of *WFDC8* pGL3[luc]promoter with the two allelic variants -44G and -44A. Luciferase activity was measured in 15 replicates.

4. Final Discussion

The study of *WFDC* and *SEMG* genes in hominids reveals contrasting evolutionary patterns in humans and chimpanzees with potential implications for both human health and chimpanzee conservation. A remarkable heterogeneity of evolutionary forces acting on these genes was detected. In humans, distinctive signals of natural selection were identified in different populations: short-term balancing selection in *WFDC8* in Europeans; incomplete selective sweep of *SPINT4* in Africans; and selection on standing variation of *SEMG1* in Asians. In chimpanzees, a single signal of strong purifying selection was detected in *WFDC6* and *EPPIN,* common to three *P. trologlodytes* subspecies. Overall, the evolution of the *WFDC* locus in humans and chimpanzees is expected to portray the differences in traits of pathogen resistance and reproductive biology in hominids.

4.1. Signals of Selection in *WFDC*s in Humans

Despite the critical contribution of the GWS in the detection of the strongest selective signals (hard sweeps) and in the understanding of the genetic basics of human adaptations, the current view of the field proposes a significant role of milder selective events (soft sweeps and selection on standing variation), as more frequent among humans and rarely captured by classical GWS (Pritchard, Pickrell and Coop 2010; Hernandez, et al. 2011). In recent studies positive selection was predicted to have targeted only 0.5% of nonsynonymous substitutions in the last 250,000 years of human evolution, positing the arising of adaptive substitutions as rare events. However, many of the recent signals of selection in humans are frequently associated with expression of quantitative trait loci (eQTLs), which include variants located in functional binding motifs of enhancers, promoters and 3'UTR (Hindorff et al., 2009; Lindblad-Toh et al., 2011; Vernot et al., 2012; Wang et al., 1995).

Most of the genes identified by the GWS for positive selection are classical targets of adaptive responses, such as skin color, lactose metabolism, and malaria resistance in which candidate variants exhibited remarkable signals (*SLC25A5*, *LCT* and *HBB*) attributable to the combination of strong of selective forces acting at very recent times (less than 50.000 years ago) (Bersaglieri, et al. 2004; Coelho, et al. 2005; McEvoy, Beleza and Shriver 2006; Kelley and Swanson 2008; Akey 2009). Nevertheless, the identification of a selective signal in less characterized genes that are not associated with a specific trait or phenotype, may turn the recognition of the

underlying mutation into a difficult task. In such instances candidate variants of selection may need to be functionally evaluated to address potential correlation between genotype and phenotype. Thus in some cases, the candidate genes gain support because they lie in functional pathways and in GO categories enriched in targets of positive selection, such as spermatogenesis and the immune response (Clark and Swanson 2005; Biswas and Akey 2006; Sabeti, et al. 2006; Blekhman, et al. 2008; Barrett and Hoekstra 2011), just like in the case of the *WFDC* locus.

The *WFDC* locus was first identified as a potential target of positive selection in humans, in the GWS performed by Voight et al (2006), where various regions were identified based on the integrated Haplotype Score (iHS) and HapMap Project data (Phase I/II). The genes that were suggested as being under positive selection in the GWS were *WFDC6* and *EPPIN* in Europeans, and *WFDC10A*, *WFDC11*, *WFDC10B*, *WFDC13* and *SPINT4* in Africans. However, this positive selection signal was not consistently detected in a number of GWS based in other statistics and/or databases. It is noteworthy that typically GWS usually report only the top 1% of the empirical distributions, which usually correspond to strongest cases of positive selection (hard sweeps) (Ronald and Akey 2005; Biswas and Akey 2006; Sabeti, et al. 2006; Sabeti, et al. 2007; Barreiro, et al. 2008; Kelley and Swanson 2008; Tennessen, Madeoy and Akey 2010). Milder selective signals, such as the ones found in the *WFDC* region within the 1 to 5% range of the empirical distribution may receive less attention (Pritchard and Di Rienzo 2010). In addition, GWS for positive selection are based in large databases of human variation (HapMap Project, Perlegen, and more recently the 1000 Genomes Project) which suffer from various degrees of ascertainment bias depending on the genotyping methodology used (SNP assignment or low-coverage sequencing). Hence potential targets of positive selection found in GWS, especially those with milder signal, require a systematic analysis of the candidate regions under selection.

The teasing apart of signals of natural selection shaping the genetic diversity at the *WFDC* locus in human populations was performed in two stages: Stage I was a sequencing study that validated selective signals proposed by the GWS scan by Voight et al (2006); in Stage II a high-throughput targeted sequencing study was performed that systematically cataloged the genomic variation across the *WFDC* locus in three human populations. The signatures of selection that were found through GWS for

positive selection by Voight et al (2006) were further explored through deep Sanger sequencing that allowed for the performance of genetic diversity analyses in the genes that showed unusual patterns of diversity (Article 1). This analysis re-centered the signal from *WFDC6* and *EPPIN* in Europeans to neighboring gene *WFDC8*, and in a similar fashion, narrowed down the candidate gene in Africa populations to *SPINT4*. In addition, the sequencing effort identified potential candidate variants of selection that would have not otherwise been identified using only GWS studies. In stage II, a Sanger sequencing strategy was used to systematically catalog all the genomic variation in coding regions (and some noncoding) and in a set of neutral evolving control regions (47 pseudogenes). The latter were used to define the demographic model that better suited the data. To this end, the model by Gutenkunst et al (2009) was selected out of a set of 7 different models that were tested. This model estimates the Out-of-Africa and European-Asian splits to have occurred 220 kya and 21.2 kya, respectively. In addition, this model also assumes a recent expansion of Europeans and Asians and migration between these two populations.

The most recent demographic scenario for human populations by Gravel et al (2011) was also used to assess the significance of outliers under Gutenkunst et al (2009). However, the two models are quite similar, and the model by Gravel et al (2011) benefits from the data generated by the “1000 Genomes Project”. In the “1000 Genomes Project” the use of next-generation DNA sequencing technologies in the analysis of thousands of samples from different population background allowed for the characterization of less frequent and younger genetic variation, which may be more informative regarding more recent demographic events. Indeed, an excess of rare variants (at a frequency ≤ 0.5%) was detected in the 1000 Genomes dataset which is likely to be due to a recent explosive growth of human populations. As mentioned previously, over the last 400 generations, the world population has expanded at least three orders of magnitude, reaching 7 billion people. Importantly, this excess of rare variants may affect the inference of positive selection signals as it implies massive departures from population genetics equilibrium potential causing a increment in the number of false positives (Jobling, Hurles and Tyler-Smith 2004; Gutenkunst, et al. 2009; Gravel, et al. 2011; Roberts 2011; Keinan and Clark 2012).

Overall these studies demonstrated that most of the *WFDC* genes evolve neutrally, showing standard patterns of genetic diversity. Still, a few exceptions demonstrated non-standard genetic diversity levels in both phases of the study: In particular, *WFDC8* showed to be under recent balancing selection in Europeans; *SPINT4* is suggested to be under an incomplete selective sweep in Africans; and *SEMG1* has demonstrated patterns of genetic variation that suggest the action of selection on standing variation in Asian populations.

Specifically in Europeans, a newly identified signal arose from *WFDC8* that presented mixed features like a long range haplotype associated with the -44A variant an atypical deep-root tree topology, which was interpreted as a signal of short-term balancing selection (Article 1). This signal was unexpected as balancing selection was thought to occur infrequently in the human genome. Moreover, it is likely to represent one of a few robust signals of balancing selection previously described in the literature. (Charlesworth 2006; Andrés, et al. 2009; Fumagalli, et al. 2009). It was hypothesized that the candidate variant -44A could potentially down regulate gene expression in the male reproductive tract, balancing the proteolytic activity as well as antiinflammatory and antibacterial activity. Furthermore, the recent balancing selective signal in Europeans surrounding *WFDC8* was further confirmed in the high-throughput sequencing study (Article 2). Taking into account these results, the signal of recent balancing selection in *WFDC8*, is indeed, the most robust signal of the three identified in the *WFDC* locus, being confirmed with remarkable concordance independently of sample scheme in CEU population. The significantly positive Tajima's *D* is confirmed by all demographic scenarios studied (Sabeti, et al. 2005; Schaffner, et al. 2005; Voight, et al. 2005; Gutenkunst, et al. 2009), including the most recent demographic model designed by Gravel (2011). Not only was the *WFDC8* signal detected by a GWS for positive selection and confirmed by both Sanger sequencing efforts, it is also supported by the European population panel from the 1000 Genomes Project (Appendix 2; Supplementary Table S5).

In Africans *SPINT4* showed unusually high values of nucleotide diversity, a significant Fay and Wu's *H* value, and a haplotype configuration [Ser73+98A] with frequencies of 80%. Taken together these data were interpreted as an incomplete selective sweep signal. Similarly to the situation of *WFDC8*, this is a new report suggesting selective pressures in this gene (Article 1). The *SPINT4* Ser73 replacement

leads to a modified Kunitz domain (of reduced inhibitory activity) and 98A maintains an active trans-activator binding site, which may increase the innate immunity function without compromising the proteolytic features (necessary for fertility) of male secretions (Article 1). In addition, despite most *WFDC* genes being under purifying selection (demonstrated by low frequency NSyn variants) the follow-up sequencing study suggested Gly73Ser as a variant under positive selection due to elevated population frequency and the most elevated F_{ST} values in Africans (Article 2). The 1000 Genomes Project database, however, does not support the nonneutral evolution of *SPINT4* in African populations, detected by the GWS (Appendix 2; Supplementary Table S5).

Interestingly, both candidate variants of selection, the -44A allele in *WFDC8* and [Ser73+98A] in *SPINT4* were proposed to confer a selective advantage through a fine regulation of their combined activities as serine protease inhibitors, anti-inflammatory, and antibacterial molecules. As these genes are not yet characterized, the interpretation of these results would benefit from a functional characterization of the candidate variants and how it affects not only the expression levels of these proteins but also their biological roles.

A number of previous studies have shown that the genes located in the *WFDC*-CEN present signals of accelerated evolution in primates and rodents (Jensen-Seaman and Li 2003; Kingan, Tatar and Rand 2003; Dorus, et al. 2004; Hurle, et al. 2007; Ramm, et al. 2008). *SEMGs*, *PI3*, *SLPI* and *WFDC5* showed high d_N/d_S values, demonstrating signals of positive selection in primates (Hurle, et al. 2007), and the evolutionary rate of *SEMG2* was positively correlated with promiscuity (Dorus, et al. 2004). Despite these signals that provide evidence of selection, none of the genes located at the *WFDC*-CEN locus displayed low empirical values in the previous GWS, most likely because the iHS statistics are designed to capture variants that have not yet reached fixation and are associated with long homogeneous haplotypes (Voight, et al. 2006).

When a high-throughput sequencing study was performed, new evidence suggested that *SEMG1* was yet another gene that did not fit the standard patterns of diversity in Asian populations. This signal had not been previously described and was associated with a skew toward low-frequency variants and significant *P*-values when corrected for demography. The selective signal was further supported by haplotype

based tests (Hudson haplotype test, DIND, and EHH/REHH) centered on the *PI3-SEMG2-SEMG1-SLPI* region. In this case, the best candidate variant of positive selection was Ser56 (rs2301366) located on the second exon of *SEMG1*. At the functional level, the Thr56Ser change is expected to impact the antimicrobial and antiviral activity of N-terminal SEMG1 peptides in both male and female reproductive tracts (Robert and Gagnon 1999; Bourgeon, et al. 2004; Edstrom, et al. 2008; Zhao, et al. 2008; Martellini, et al. 2009). However, the occurrence of Thr56Ser at an 80% frequency in Asians and in other populations at intermediate frequencies, together with the fact that *SEMG1* was not detected in the GWS of Voight et al (2006), lead to the proposal of *SEMG1* being an example of a gene under selection on standing variation (soft sweep) (Przeworski, Coop and Wall 2005; Pritchard and Di Rienzo 2010; Hernandez, et al. 2011). Conversely, the *SEMG1* signal in Asians presents a significantly low Tajima's *D* in the 1000 Genomes Project database, even when sample size was increased (Appendix 2; Supplementary Table S5).

Importantly, selection on standing variation may underlie *SEMG1* and *SPINT4* signals which, depending on the value of frequency of the advantageous variant at the time of the selective event, can lead to a large variance in allele frequency spectra and in haplotype patterns leading to confounding and complex selective signals. Such results are in agreement with the reported findings of soft sweeps being more common in humans. However these signals are still largely undetected when using the classical statistical methods and require a candidate gene centered approach, as complex pleiotropic selective signals are hardly captured by classical GWS (Pritchard and Di Rienzo 2010; Pritchard, Pickrell and Coop 2010; Hernandez, et al. 2011).

The 1000 Genomes Project made publicly available a significant amount of samples that represent 21 human populations to a total of 1092 low coverage genomes and exome sequences with 50x coverage. Even though they do not yet completely replace deeper coverage and high-quality Sanger sequencing studies, this data can be of great use to survey the genomic diversity of human populations, to find potential outliers and to select a subsample dataset that could be sequenced and analyzed to confirm results. The 1000 Genomes Project database underrepresents low-frequency variants in the low coverage regions, which could be a disadvantage when trying to detect signals of positive selection. A signature of selection that modifies the site frequency spectrum (for example, resulting in an increase of singletons and

doubletons) would likely remain undetected if only relying on the genetic variability available in this database. Nonetheless, the coherence of two of the three uncovered signals through high-throughput Sanger sequencing combined with the 1000 Genomes database further highlights that *WFDC8* and *SEMG1* are genes evolving under non-neutral pressures.

WFDC selective signals suggest heterogeneous responses to geographical environmental changes with very population specific signals that can hypothetically be correlated with the large heterogeneity of pathogenic agents and infectious burden across the globe (Novembre and Di Rienzo 2009; Fumagalli, et al. 2011; Hancock, et al. 2011). However, only future studies with more detailed information about pathogen diversity combined with a detailed study of antimicrobial properties of candidate variants may grasp the environment-gene interaction that has shaped *WFDC8*, *SEMG1* and *SPINT4*. Besides, other unknown biological function(s) of the *WFDC* genes cannot be discarded as selective driven forces of *WFDC*s, like for instance, a regulation of proteolysis in the reproductive track or in other tissues.

On the other hand, additional relevant information can be obtained from the *WFDC* genetic variation in ethnically diverse populations to address the extent and the strength of selective pressures acting in this locus. Various studies argue that biological mechanisms underlying a given adaptive phenotype may vary across populations, resulting from the same or different variants. For example, mutations at the *FUT2* gene differ among Eurasian and African populations, demonstrating a complex picture which points toward natural selection and its different actions stratified according to geography (Ferrer-Admetlla, et al. 2009). Similarly, in the *G6PD* gene a different variant from the one observed in Africans is found in Southeast Asians, reducing *Plasmodium vivax* but not *P. falciparum* infection and indicating the first malaria pathogen as the major driving force behind the strong selective signal for *G6PD* in Southeast Asians (Louicharoen, et al. 2009; Hedrick 2011). In addition, the newly available genome datasets representing various populations have revived interest in the study of geographic correlations between allele frequencies and environmental variables. Various studies suggest that genes that are potential susceptibility loci in common human diseases have marked inter-ethnic differences in prevalence. An example is the correlation of *AGT* and *CYP3A* variants with the higher prevalence of hypertension, and in particular salt-sensitive hypertension, in African Americans

compared with European Americans, reflecting adaptations to hot equatorial climates in ancestral African populations (Thompson, et al. 2004; Coop, et al. 2009; Novembre and Di Rienzo 2009).

Overall, a deeper understanding of the evolutionary processes that generate genetic differences between populations, and an improved knowledge on how natural selection affects the different human populations will hopefully provide spatial patterns of genetic variation and help to determine better ways to prevent diseases in a particular human population.

4.2 Signals of Selection of *WFDC*s in Chimpanzees

To better describe the evolution of *WFDC*s and characterize the extent of genetic diversity shared within and between hominoids, a systematic comparative genomics and population genetics analysis was performed utilizing high-throughput sequencing. The *WFDC* locus and the same 47 control regions were Sanger sequenced in three chimpanzee subspecies (*P. t. troglodytes*, *P. t. verus* and *P. t. ellioti*) and were contrasted in terms of polymorphism and genetic variation. In addition to studying the genetic variability in *WFDC* genes, the description of the subspecies of the sequenced individuals was also of interest.

The common chimpanzee, *Pan troglodytes,* and bonobos, *P. paniscus,* are two different species currently habiting in West and Central Africa, separated by the geographical barrier of the Congo River (Figure 7). Common chimpanzees are further divided into four subspecies: *P. t. verus*, located in western Africa and occupying the Upper Guinea region; *P. t. troglodytes* extending throughout central Africa; *P. t. schweinfurthii* living in eastern Africa; and *P. t. ellioti*, that occurs in the Gulf of Guinea (Nigeria and Cameroon) in a region limited by the Niger and Sanaga rivers (Figure 7).

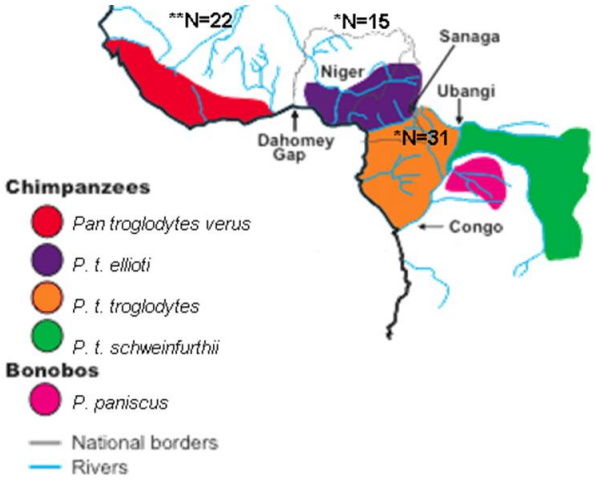


Figure 7: Geographic distribution of *Pan troglodytes* and *P. paniscus* in Africa (adapted from Gonder et al. 2011).

Although the demographic history of chimpanzees is described in less detail than humans, several studies support an elevated effective population size for *P. troglodytes* around 0.50 million years ago. Furthermore, *P. t. verus* is known to share a common ancestor with the remaining subspecies approximately 0.46 mya, and *P. t. troglodytes* and *P. t. schweinfurthii* to have diverged 0.10 mya. *P. t. ellioti* was proposed to have separated around 0.32 mya. At the genetic level *P. t. verus* shows the most striking differences between subspecies; *P. t. verus* are currently isolated from other chimpanzees and undergone strong population bottlenecks, which had eliminated most of the pre-existing diversity of this subspecies. Consequently, nucleotide diversity in these western chimpanzees reached low levels such as that described for humans. Furthermore, hybridization seems to occur occasionally in the wild between *P. t. ellioti* and *P. t. troglodytes*, although they are thought to be genetically isolated from each other (Gonder, Disotell and Oates 2006; Stone, et al. 2010; Gonder, et al. 2011; Bowden, et al. 2012).

Accordingly, when the *WFDC* region was analyzed there was a need to circumvent the lack of demographic information regarding the three subspecies sampled (Article 3). To that end, the control regions (neutrally evolving regions) were used in an attempt to describe the genetic variation in these three populations and to the state of knowledge on the differentiation of *P. t. ellioti*. In contrast to other studies, the results shown were not conclusive when considering *P. t. ellioti* as a genetically isolated subspecies. Previous studies have used the same set of individuals, but it

appears that the markers used here are not sufficient to discriminate the two closely related subspecies (*P. t. ellioti* and *P. t. troglodytes*), even with the addition of markers from the *WFDC* locus to a total of 1268 SNP markers. This finding suggests that the markers used in this study are not the recommended type to determine population structure and differentiation. Most of the studies that successfully describe *P. t. ellioti* as a separate subspecies used markers with high mutation rates (mtDNA and microsatellites), which contain more information about genetic diversity between populations (Gonder, et al. 2011; Bowden, et al. 2012). Moreover, when using autosomal SNP markers such as the ones in this study, additional markers should be genotyped and further confirmation of individual relationship with mtDNA should be undertaken (Bowden, et al. 2012).

Despite the aforementioned results, the empirical distribution built from the control regions did lead to the discovery of a signature of selection in *WFDC6* and *EPPIN* in *P. t. troglodytes:* two paralog *WFDC* genes under purifying selection. The negative Tajima's *D* values obtained for these two genes were compared with the lower percentile of empirical distribution built with neutral regions. Coalescent simulations corrected for demography of *P. t. troglodytes* also supported a significant result for *WFDC6* but not for *EPPIN*. In the case of *WFDC6*, strong purifying selection seems the most probable basis for the departure of neutrality. This hypothesis is sustained by the low levels of population differentiation present in all sampled subspecies and the absence of a reliable candidate variant of positive selection. Additionally, the d_N/d_S ratios of *WFDC6* and *EPPIN* (71% similarity) suggest a rapid evolution of *WFDC6* after gene duplication, followed by a period of strong functional constraints. Both *WFDC6* and *EPPIN* present evidence of purifying selection, potentially due to their roles in reproduction and immunity in primates.

WFDC6 and *EPPIN* are recent parologs which share the same WAP and Kunitz domains, and conserved most of the disulfide bridges crucial to the maintenance of WFDC three dimensional structures, antimicrobial and antiviral activities, immunity, and tumorogenic features. However, *WFDC6* diverges from *EPPIN* in the composition of the amino acid located at the reactive site, with several substitutions pointing to a probable loss of PSA inhibitory activity. Due to chimpanzees' multimale/multifemale mating systems and promiscuity, they are more likely to be exposed to sexually transmitted pathogens in which the conservation of an additional antimicrobial protein,

like WFDC6, would be advantageous. Disease transmission during mating provides a connection between reproduction and immunity, where sexually transmitted diseases can affect fitness of individuals by imposing different selective pressures on their hosts. Hence, the pseudogenization of *WFDC6* observed in other primate species (orangutans and baboons) may be correlated with a lower pathogen load in unimale mating systems or through different immunological mechanisms linked to non-*WFDC* genes.

Consistent with the evolution of *WFDC6*, other genes involved in reproduction have been correlated with multimale/multifemale mating systems and female promiscuity. This is the case in *SEMG2*, which was found to have undergone an accelerated evolution in chimpanzees. However, no evidence of selection was detected at the intraspecific level for *SEMG2* (Article 3). This result suggests that event of positive selection in *SEMG2* is ancestral to chimpanzee diversification and it is only measurable through multispecies comparisons.

In the particular case of chimpanzees, it is possible that the large effective population size caused the identification of recent selective events to be more difficult. In fact, the evidence for recent positive selection in chimpanzee populations is reduced to a few examples that could alternatively be explained by the low awareness of the evolutionary forces affecting the site frequency spectrum in chimpanzees. Only recently have efforts been made to generate large genetic variation data that may allow further developments in this field of research (Wegmann and Excoffier 2010; Hai Yang Hu1. and Liang3 2011; Hobolth, et al. 2011; Auton, et al. 2012; Bowden, et al. 2012; Scally, et al. 2012). Here, the public database PanMap may have an important contribution as it allowed for completion of the first chimpanzee recombination map [PanMap; (Auton, et al. 2012)]. However, PanMap relies only on *P. t. verus* data, which is the most homogeneous chimpanzee subspecies.

Not surprisingly, the evolution of other genes involved in reproduction has been correlated with multimale/multifemale mating systems and promiscuity such as *SEMG2*, which is under rapid evolution in chimpanzees. However, this signal is an older positive selection event that was detected by multispecies comparisons but, it was not consistent at the intraspecific level in *SEMG2* (Article 3). The evolutionary

study of the *WFDC* genes would have benefited from a deeper understanding of how the chimpanzee genome and populations evolved until the present date.

4.3 *WFDC*s in Hominids

Even though humans and chimpanzees are two closely related species, their mating systems and immune responses are highly diverged. Humans occupy most geographical areas of the world and thus, are exposed to a wide range of different environments and consequently, are subjected to very different selective pressures. Accordingly, the *WFDC* genes appear to have been targeted by more recent selective events in humans than chimpanzees, which might be correlated with human migrations to different geographical regions. In humans, the signals of positive selection are population specific, centered in *WFDC8*, *SPINT4* and *SEMG1*, in Europeans, Africans, and Asians respectively (Article 1 and 2). Conversely, in chimpanzees which have been confined to the African tropical area, the signal of purifying selection in *WFDC6* is likely to affect the entire species equally (Article 3).

The heterogeneity of selective signals observed among *WFDC* genes in humans suggests the occurrence of multiple adaptive hits in different environments, possibly correlated with the presence or absence of different pathogens per geographical region (Prugnolle, et al. 2005; Balaresque, Ballereau and Jobling 2007; Bhushan, et al. 2009; Fumagalli, et al. 2009; Wlasiuk and Nachman 2010a; Fumagalli, et al. 2011; Seixas, et al. 2011). On the other hand, chimpanzees seem to conserve *WFDC6* and *EPPIN*, in order to potentially target a larger spectrum of proteases and to increase antimicrobial potential in a possible response to the selective pressures caused by STDs (Chimpanzee Sequencing and Analysis Consortium 2005). Previous studies have demonstrated the direct correlation between indicators of immunocompetence (such as increased levels of leukocytes) and proxies of female sexual promiscuity among primates with different mating systems, however no correlation with other social and ecological variables was reported (Nunn, Gittleman and Antonovics 2000; Nunn 2002b, 2002a; Nunn, et al. 2003; Holmes 2004; Wlasiuk and Nachman 2010a, 2010b; Garamszegi and Nunn 2011). This further emphasizes that promiscuous species like chimpanzees, as opposed to humans, are more exposed to STDs, which has induced a greater investment in genes enrolled in immune response (Wlasiuk and Nachman 2010a, 2010b). Biologically, the absence of

replication in the signals between humans and chimpanzees suggests major selective forces to be driven by different exposures to pathogens.

Due to the intimate relationship between evolutionary rates of *SEMG*s, the signals of *SEMG*s and *WFDC*s in humans and chimpanzees were found to be uncorrelated. The *SEMG*s have an active role in reproductive biology and are thought to be under non-neutral evolution, nevertheless *SEMG*s behave differently in humans than in chimpanzees. *SEMG1* which shows a signal of recent selection in Asians associated with an NSyn variant (Ser56); has been shown in primates to be under rapid evolution, partially due to a rapid expansion of a repetitive stretch which enables the crosslinking of SEMG proteins in the semen coagulum (Jensen-Seaman and Li 2003; Kingan, Tatar and Rand 2003; Jobling, Hurles and Tyler-Smith 2004; Carnahan and Jensen-Seaman 2008; Ramm, et al. 2008). However, the evolutionary rate of *SEMG1* does not seem to be correlated with promiscuity levels in primates (Hurle, et al. 2007). Conversely, *SEMG2* has demonstrated signals of adaptive evolution in primates by demonstrating a correlation between NSyn substitutions and the mean of male partners per periovulatory period of a female (Dorus, et al. 2004). The evolution of *SEMG2* is accelerated in polyandrous primates relative to monandrous primates, showcasing the intimate relationship between sexual selection and the molecular evolution of reproductive genes. Despite the evidence for selection targeting *WFDC* and *SEMG*s genes due to their roles in reproduction and host-pathogen interaction, the genes under selection in humans and chimpanzees do not overlap. *SEMG1* is the only exception, even though the selective forces driving the evolution of this gene are not the same. In chimpanzees, *SEMG1* evolution appears to be driven by its role in post copulatory selection in order to thicken the semen coagulum through increased crosslinking events, even though d_N/d_S ratios do not support this observation (de Lamirande, et al. 2001; Jensen-Seaman and Li 2003; Kingan, Tatar and Rand 2003; de Lamirande 2007; Malm, et al. 2007). Conversely, in humans, the *SEMG1* signal of selection is probably due to host-pathogen interactions in the reproductive tract in Asian populations.

Around 5.4 million years ago humans and chimpanzees started to diverge (Patterson, et al. 2006), which subjected them to various environment changes and population history leaving signatures of their evolutionary history at the genomic level. In general, chimpanzees exhibit higher levels of differentiation between subspecies due

to their large effective size, than humans between populations. The only exceptions to this global pattern are *WFDC6* and *EPPIN* which, due to strong purifying selection in chimpanzees, present nucleotide diversity levels in the same order of magnitude as that observed in humans.

The extension of LD blocks in the *WFDC* locus largely differs between humans and chimpanzees and while in humans varies from 10 kb to 100 kb, on average; in chimpanzee LD blocks are nearly absent or encompass a few kb. Again, this could be due to a higher effective population size of chimpanzees and to the fact that recombination hotspots are not shared between these two species (Ptak, et al. 2005; Fledel-Alon, et al. 2009; Baudat, et al. 2010; Fledel-Alon, et al. 2011; Segurel, Leffler and Przeworski 2011). Even though the recombination rates among humans and chimpanzees do not significantly differ, haplotype blocks in chimpanzees are not localized in the same regions and the nucleotide diversity levels are a significantly higher in chimpanzees (Ptak, et al. 2004; Ptak, Voelpel and Przeworski 2004; Frazer, et al. 2007; Auton, et al. 2012). In fact, this may have had an impact on the detection of signals of positive selection in chimpanzees. The higher nucleotide diversity and lack of long homogeneous haplotypes may have influenced the ability to detect recent selective sweeps as most of the statistical tests and background evolutionary history are human centric.

4.4 Biological relevance of genes under selection

Most of the *WFDC* genes with known biological functions have combined activities in immunity and reproduction. For cases like *WFDC8*, to which no specific biological function has been attributed, it is important to follow up on the biological significance of candidate genes. A number of GWS for positive selection and Genome-Wide Association Studies (GWAS) have pinpointed candidate genes and genomic regions for future study, which include genetic factors affecting insulin resistance, cancer susceptibility, immunity disorders and others. However, these studies only nominate candidates, and it takes biological insight and studies of molecular pathways to learn how they affect human health. Even for genes with well described roles and recognized molecular pathways, it remains hard to predict the effects of common variants (Dermitzakis and Clark 2009). With the release of the 1000 Genomes Project

and the efforts to accurately characterize copy number variation, the methodology for identifying genetic differences between cases and controls is heading toward a golden opportunity for making substantial progress in unraveling disease susceptibility and advancing a deeper understanding of the biological significance of genetic variation (Tillier and Charlebois 2009; The 1000 Genomes Project Consortium 2010; Mills, et al. 2011). Despite the current release of data for even rarer variants in very large and well-designed cohort samples, functional characterization of candidate genes should be persued to better understand the effects of individual genes in tissue orientated networks.

One of the main motivations of this comprehensive study of the *WFDC* locus was to prioritize the most interesting genes and genetic variants for biological characterization to test the biological impact of the candidate genes under positive selection. Preliminary characterization was focused on *WFDC8*, as this gene demonstrated a very robust and unexpected balancing selection signal that was very striking even when compared with other genes under balancing selection in the literature (Charlesworth 2006; Andrés, et al. 2009; Fumagalli, et al. 2009; Andrés, et al. 2010). In addition, the location of the candidate variant -44A in a putative regulatory element made the experimental design relatively straight forward using a luciferase reporter assay. In addition, by undertaking molecular biology and expression assays *in vivo* the potential consequences of the selective signal found in European populations were assessed.

WFDC8 was shown to display a signature of balancing selection in Europeans associated with a -44(G/A) polymorphism (rs7273669), located in the 5' region of *WFDC8*, which was predicted to alter gene expression by abolishing the binding site of two transcription factors (Article 1 and 2). In a recent study, the rat homolog, *Wfdc8* was characterized and identified as an androgen dependent gene induced by LPS, and at the protein level, Wfdc8 was found to display a potent antibacterial activity (Rajesh, Madhubabu and Yenugu 2011). Furthermore, comparative analysis of orthologs in twelve mammalian lineages, demonstrates that *WFDC8* preserves from rodents to humans three WAP and one Kunitz domains, which are expected to combine antimicrobial and antiprotease properties (Bingle and Vyakarnam 2008). However, some residues are likely to have been targeted by positive selection, as is the case of E139, located in the reactive center of the Kunitz domain, which might have evolved

faster in *WFDC8* to counteract the activity of different proteases. Conversely, the maintenance of the 15 disulfide bridges in WFDC8 indicates a possible conservation of antimicrobial function in most mammals.

To determine the effect of the candidate variant -44(G/A), a series of expression assays were performed. Although the antimicrobial or anti-protease properties of WFDC8 were not confirmed, its expression was found to be higher in testis and lower in ovaries, bladder, and brain tissues in a panel of 19 tissues tested. Moreover, the effect of the candidate variant -44G/A in *WFDC8* expression was also investigated *in vivo*, by the genotyping of a surrogate marker in complete LD with -44(G/A) and *in vitro* by performing luciferase expression assays in two different cell lines. Preliminary results gathered from *in vivo* and *in vitro* experiments were not consistent: Specifically, while the *in vivo* assays carried out in ovaries suggest a threefold expression increase of the -44A allele, *in vitro* results suggest -44A increased expression in HeLa but not in Caco-2 cell lines.

The effect of -44G/A polymorphism in *WFDC8* expression appears to depend on the cellular context and thus balancing selection may be acting to a fine tuned regulation of gene expression, in which being a heterozygous may be advantageous. Importantly, WFDC8 protein was already identified in semen (Batruch, et al. 2011; Kagedan, et al. 2012) and the loss of its gene expression was associated with a fertility impairment after vas deferent renamastosis (Batruch, et al. 2011) and in Hutterite populations *WFDC8* appears to be associated with transmission distortion (Meyer, et al. 2012).

The identification of disease genes and variants is not meaningful unless biological processes leading to a phenotype are followed up by a deep functional characterization. Furthermore, biological processes underlying a disease phenotype may vary among tissues, where different molecular backgrounds can impact the resulting cellular outcome, as demonstrated by the different effect of the *WFDC8* candidate variant in two different cell lines. Therefore, inference of gene regulatory networks and biological context are imperative to make better educated guesses on candidate genes studies.

4.5 Implications in Human health

Most of the genes in the *WFDC* locus have not been associated with repercussions in human health, but a few studies regarding these genes have pointed out potential deleterious effects of common variants in fertility and disease susceptibility. For instance, in vivo, eppin-semenogelin binding is a critical step in the removal of semenogelin during semen liquefaction, making eppin a target for male contraception (O'Rand, et al. 2009). Other studies have examined the effect of anti-eppin antibodies in fertile male monkeys and the effect of recombinant human semenogelin on human sperm motility. In human semen, anti-eppin antibodies significantly decreased the progressive motility of spermatozoa by decreasing total distance traveled, straight-line distance, and velocity (Richardson, et al. 2001; O'Rand, et al. 2004; Wang, et al. 2007a; O'Rand, et al. 2011; O'Rand, et al. 2011; O'Rand and Widgren 2012).

In a study of human male infertility centered in *EPPIN*, two variants were associated with different male infertility phenotypes. The variant rs2231829 located 5' of the *EPPIN* transcription start was associated with a significant decreased risk of idiopathic infertility, and the variant rs11594 located in *EPPIN*'s fourth exon was associated with an increased risk of idiopathic infertility and abnormal semen parameters (Ding, et al. 2010). Conversely, mass spectrometry studies have identified SEMG2 as one of the few proteins with differences in semen protein profiles of fertile and infertile men; where SEMG2 was nearly absent in the semen of infertile men (Thacker, et al. 2011). In addition, seminal plasma motility inhibitors (SPMIs) resulting from the proteolytic cleavage of SEMG1 and SEMG2, are thought to be correlated with sperm motility, as they appear to remain on the surface of sperm after liquefaction. This might account for some disorders of sperm motility observed in infertile men with asthenozoospermia (Terai, et al. 2010).

EPPIN and *SEMGs* implications for human health are not restricted to infertility. The effects of *SEMGs* overexpression on prostate cancer and cell proliferation suggest that they can be used in the prognosis of cancer progression after radical prostatectomy (Canacci, et al. 2011). Secreted *SEMG1*, *SEMG2*, and *EPPIN* were all significantly lower in carcinoma patients than in individuals with benign or prostatic intraepithelial neoplasia (Izumi, Zheng and Miyamoto 2011; Izumi, et al. 2012).

However, the biological significance of *SEMGs* overexpression in prostate cancer still remains to be fully developed (Canacci, et al. 2011). Furthermore, EPPIN protein complex components in sperm cells including SEMG1, one isoform of CLU, and two isoforms of LTF, were found to be enhanced in type-1 and type-2 diabetic individuals. The correlation is not yet deeply understood, but it could possibly results from a response to enduring hyperglycemia and enhanced oxidative stress (Paasch, et al. 2011).

Semen is known to be the main vehicle of transmission of the HIV virus, with most of the studies showing that SEMG1-derived peptides have a protective anti-HIV activity (Martellini, et al. 2009). One recent study has, however, questioned the anti-HIV properties of cationic peptides resulting from the SEMG1 and SEMG2 cleavage, in which a subset of SEMG-derived fragments appear to greatly enhance HIV infection (Roan, et al. 2011). On the other hand, the antiviral activity of *SEMG* peptides was demonstrated by one peptide fragment of SEMG1 that was purified from seminal plasma. The peptide showed a transient anti-HIV activity that, after prolonged exposure to proteases was degraded, losing their antiviral activity (Martellini, et al. 2009). In contrast to semen samples from healthy individuals that greatly enhance HIV infection, SEMGs-deficient semen samples from patients with ejaculatory duct obstruction do not present enhanced HIV activity (Roan, et al. 2011). In lieu of these controversial observations, the role of SEMG-derived peptides in immune response require additional clarification and the characterization of the Thr56Ser variant, associated with a selection signal, should also be pursued.

The examples of *SEMGs* and *EPPIN* implications in human health are valuable approximations for the roles of genes under selection such as *WFDC8* and *SPINT4* in infertility, immune response and even complex disorders like diabetes. The importance of following up and determining the biological significance of candidate genes under selection cannot be minimized in the quest to describe the pathways they are involved in, and the repercussions of these genes in disease susceptibility.

5. Conclusions

This work lends support to the notion that genes related to host-pathogen interactions and reproductive biology, such as the *WFDC gene*s, are usual targets of natural selection.

It highlights and elucidates signals of natural selection that have shaped the genetic diversity at the *WFDC* locus in human populations. Three different *WFDC* genes are evolving under selection in three different human populations, suggesting these genes are under different selective pressures and are responding to different pathogen loads in various geographic areas (Article 1 and Article 2). These signals are as follows:

A signal of recent balancing selection centered in *WFDC8* in Europeans and associated to the variant -44(G/A) in the 5' gene region, was newly identified and demonstrated to be the most robust signal as it was detected in Stage I and II of the sequencing studies and further confirmed in the 1000 Genomes Project database.

In African populations, an incomplete selective sweep signal was discovered around *SPINT4*, where the haplotypic configuration Ser73+98A were determined to be the most probable candidate variants.

In Asian populations, a new signal of selection around *SEMG1* was detected in Stage II. *SEMG1* appears to be under selection on standing variation, where a pre-existing variant Thr56Ser reached high frequency. In addition, this signal was confirmed in the 1000 Genomes Project database.

This study highlights that genome-wide scans for natural selection are indeed designed to detect mostly hard sweeps. In the wider context, it confirms that soft sweeps and polygenic adaptation, which may actually be more common than classical selective sweeps, are possibly being missed by GWS studies.

This work also demonstrates that the genetic diversity levels at the *WFDC* locus within and between hominids are divergent. No overlap in selective footprints was detected at the *WFDC* locus between humans and chimpanzees (Article 3).

A plausible hypothesis is that the gene *WFDC6*, a paralog of *EPPIN*, may have evolved in order to provide additional response to pathogen infections and is being maintained divergent from *EPPIN* by purifying selection. Given that chimpanzees are a promiscuous species, a possibility is that they are subjected to more sexually transmitted diseases where the rapid evolution of *WFDC6* may be a response to such selective pressure.

Furthermore, the characterization of the demographic history of the sampled chimpanzee subspecies (*P. t. troglodytes*, *P. t. ellioti* and *P. t. verus*) by using high quality sequence data and approximately one thousand markers was unsuccessful. The choice of the genetic markers is an important aspect in studies of demographic history and different markers with higher mutation rates should be added in future research to help reach this goal.

It provides preliminary results on the biological impact of one of the candidate genes, *WFDC8*, under recent balancing selection. The balancing selection signal of *WFDC8* was confirmed in multiple ways, and the additional variation of this gene may provide a more timely response to environmental and cellular background changes (Article 4). Highly conserved in 12 mammalian species, *WFDC8* probably maintains its antimicrobial role, as its rat homolog does. Although preliminary *in vitro* and *in vivo* studies suggest a functional impact of the candidate variant -44A in the expression and regulation of *WFDC8*, an improved experimental design will be needed to clarify this issue.

6. References

Aagaard JE, Yi X, MacCoss MJ, Swanson WJ 2006. Rapidly evolving zona pellucida domain proteins are a major component of the vitelline envelope of abalone eggs. Proceedings of the National Academy of Sciences of the United States of America 103: 17302-17307.

Abi-Rached L, Moesta AK, Rajalingam R, Guethlein LA, Parham P 2010. Human-specific evolution and adaptation led to major qualitative differences in the variable receptors of human and chimpanzee natural killer cells. PLoS Genetics 6: e1001192.

Akey JM 2009. Constructing genomic maps of positive selection in humans: where do we go from here? Genome Res 19: 711-722.

Akey JM, Eberle MA, Rieder MJ, Carlson CS, Shriver MD, Nickerson DA, Kruglyak L 2004. Population History and Natural Selection Shape Patterns of Genetic Variation in 132 Genes. PLoS biology 2.

Altshuler DM, Gibbs RA, Peltonen L*, et al.* 2010. Integrating common and rare genetic variation in diverse human populations. Nature 467: 52-58.

Andrés AM, Dennis MY, Kretzschmar WW*, et al.* 2010. Balancing selection maintains a form of ERAP2 that undergoes nonsense-mediated decay and affects antigen presentation. PLoS Genetics 6: e1001157.

Andrés AM, Hubisz MJ, Indap A*, et al.* 2009. Targets of balancing selection in the human genome. Molecular Biology and Evolution 26: 2755-2764.

Auton A, Fledel-Alon A, Pfeifer S*, et al.* 2012. A fine-scale chimpanzee genetic map from population sequencing. Science 336: 193-198.

Balaresque PL, Ballereau SJ, Jobling Ma 2007. Challenges in human genetic diversity: demographic history and adaptation. Human molecular genetics 16 Spec No: R134-139.

Barreiro LB, Ben-Ali M, Quach H*, et al.* 2009. Evolutionary dynamics of human Toll-like receptors and their different contributions to host defense. PLoS Genetics 5: e1000562.

Barreiro LB, Laval G, Quach H, Patin E, Quintana-Murci L 2008. Natural selection has driven population differentiation in modern humans. Nature genetics 40: 340-345.

Barreiro LB, Quintana-Murci L 2010. From evolutionary genetics to human immunology: how selection shapes host defence genes. Nature reviews. Genetics 11: 17-30.

Barrett RD, Hoekstra HE 2011. Molecular spandrels: tests of adaptation at the genetic level. Nat Rev Genet 12: 767-780.

Batruch I, Lecker I, Kagedan D, Smith CR, Mullen BJ, Grober E, Lo KC, Diamandis EP, Jarvi KA 2011. Proteomic analysis of seminal plasma from normal volunteers and post-vasectomy patients identifies over 2000 proteins and candidate biomarkers of the urogenital system. J Proteome Res 10: 941-953.

Baudat F, Buard J, Grey C, Fledel-Alon A, Ober C, Przeworski M, Coop G, de Massy B 2010. PRDM9 is a major determinant of meiotic recombination hotspots in humans and mice. Science 327: 836-840.

Bersaglieri T, Sabeti PC, Patterson N, Vanderploeg T, Schaffner SF, Drake JA, Rhodes M, Reich DE, Hirschhorn JN 2004. Genetic Signatures of Strong Recent Positive Selection at the Lactase Gene. Am. J. Hum. Genet. 74: 9.

Bhushan S, Schuppe HC, Fijak M, Meinhardt A 2009. Testicular infection: microorganisms, clinical implications and host-pathogen interaction. Journal of reproductive immunology 83: 164-167.

Bingle CD, Vyakarnam A 2008. Novel innate immune functions of the whey acidic protein family. Trends in immunology 29: 444-453.

Bingle L, Cross SS, High AS, Wallace WA, Rassl D, Yuan G, Hellstrom I, Campos MA, Bingle CD 2006. WFDC2 (HE4): a potential role in the innate immunity of the oral cavity and respiratory tract and the development of adenocarcinomas of the lung. Respiratory research 7: 61.

Bingle L, Singleton V, Bingle CD 2002. The putative ovarian tumour marker gene HE4 (WFDC2), is expressed in normal tissues and undergoes complex alternative splicing to yield multiple protein isoforms. Oncogene 21: 2768-2773.

Biswas S, Akey JM 2006. Genomic insights into positive selection. Trends in genetics : TIG 22: 437-446.

Blekhman R, Man O, Herrmann L, Boyko AR, Indap A, Kosiol C, Bustamante CD, Teshima KM, Przeworski M 2008. Natural selection on genes that underlie human disease susceptibility. Current biology : CB 18: 883-889.

Bouchard D, Morisset D, Bourbonnais Y, Tremblay GM 2006. Proteins with whey-acidic-protein motifs and cancer. The Lancet Oncology 7: 167-174.

Bourgeon F, Evrard B, Brillard-Bourdet M, Colleu D, Jegou B, Pineau C 2004. Involvement of semenogelin-derived peptides in the antibacterial activity of human seminal plasma. Biology of reproduction 70: 768-774.

Bowden R, MacFie TS, Myers S, Hellenthal G, Nerrienet E, Bontrop RE, Freeman C, Donnelly P, Mundy NI 2012. Genomic Tools for Evolution and Conservation in the

Chimpanzee: Pan troglodytes ellioti Is a Genetically Distinct Population. PLoS Genetics 8: e1002504.

Bronson PG, Mack SJ, Erlich HA, Slatkin M 2013. A sequence-based approach demonstrates that balancing selection in classical human leukocyte antigen (HLA) loci is asymmetric. Hum Mol Genet 22: 252-261.

Bustamante CD, Fledel-Alon A, Williamson S*, et al.* 2005. Natural selection on protein-coding genes in the human genome. Nature 437: 1153-1157.

Bustamante CD, Ramachandran S 2009. Evaluating signatures of sex-specific processes in the human genome. Nature genetics 41: 8-10.

Cagliani R, Fumagalli M, Riva S, Pozzoli U, Comi GP, Menozzi G, Bresolin N, Sironi M 2008. The signature of long-standing balancing selection at the human defensin beta-1 promoter. Genome Biology 9: R143.

Cai Z, Tchou-Wong K-M, Rom WN 2011. NF-kappaB in Lung Tumorigenesis. Cancers 3: 4258-4268.

Campbell MC, Ranciaro A, Froment A*, et al.* 2012. Evolution of functionally diverse alleles associated with PTC bitter taste sensitivity in Africa. Molecular Biology and Evolution 29: 1141-1153.

Canacci AM, Izumi K, Zheng Y, Gordetsky J, Yao JL, Miyamoto H 2011. Expression of semenogelins I and II and its prognostic significance in human prostate cancer. Prostate 71: 1108-1114.

Carnahan SJ, Jensen-Seaman MI 2008. Hominoid seminal protein evolution and ancestral mating behavior. American journal of primatology 70: 939-948.

Charlesworth D 2006. Balancing Selection and Its Effects on Sequences in Nearby Genome Regions. PLoS Genetics 2.

Chhikara N, Saraswat M, Tomar AK, Dey S, Singh S, Yadav S 2012. Human epididymis protein-4 (HE-4): a novel cross-class protease inhibitor. PLoS One 7: e47672.

Chimpanzee Sequencing and Analysis Consortium 2005. Initial sequence of the chimpanzee genome and comparison with the human genome. Nature 437: 69-87.

Clark AG 2006. Genomics of the evolutionary process. Trends in ecology & evolution 21: 316-321.

Clark NL, Aagaard JE, Swanson WJ 2006. Evolution of reproductive proteins from animals and plants. Reproduction 131: 11-22.

Clark NL, Gasper J, Sekino M, Springer SA, Aquadro CF, Swanson WJ 2009. Coevolution of interacting fertilization proteins. PLoS Genetics 5: e1000570.

Clark NL, Swanson WJ 2005. Pervasive adaptive evolution in primate seminal proteins. PLoS Genetics 1: e35.

Clauss A, Lilja H, Lundwall A 2005. The evolution of a genetic locus encoding small serine proteinase inhibitors. Biochemical and biophysical research communications 333: 383-389.

Clauss A, Lilja H, Lundwall A 2002. A locus on human chromosome 20 contains several genes expressing protease inhibitor domains with homology to whey acidic protein. Biochem J. 368: 9.

Clauss A, Persson M, Lilja H, Lundwall A 2011. Three genes expressing Kunitz domains in the epididymis are related to genes of WFDC-type protease inhibitors and semen coagulum proteins in spite of lacking similarity between their protein products. BMC biochemistry 12: 55.

Coelho M, Luiselli D, Bertorelle G, Lopes AI, Seixas S, Destro-Bisol G, Rocha J 2005. Microsatellite variation and evolution of human lactase persistence. Hum Genet 117: 10.

Coop G, Pickrell JK, Novembre J*, et al.* 2009. The Role of Geography in Human Adaptation. PLoS Genet 5.

Crisci JL, Wong A, Good JM, Jensen JD 2011. On characterizing adaptive events unique to modern humans. Genome Biol Evol 3: 791-798.

Darwin C. 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life: J. Murray.

de Lamirande E 2007. Semenogelin, the main protein of the human semen coagulum, regulates sperm function. Seminars in thrombosis and hemostasis 33: 60-68.

de Lamirande E, Yoshida K, Yoshiike TM, Iwamoto T, Gagnon C 2001. Semenogelin, the main protein of semen coagulum, inhibits human sperm capacitation by interfering with the superoxide anion generated during this process. J Androl 22: 672-679.

Dermitzakis ET, Clark AG 2009. Genetics. Life after GWA studies. Science 326: 239-240.

Ding X, Zhang J, Fei J*, et al.* 2010. Variants of the EPPIN gene affect the risk of idiopathic male infertility in the Han-Chinese population. Human reproduction 25: 1657-1665.

Dixson AF, Anderson MJ 2004. Sexual behavior, reproductive physiology and sperm competition in male mammals. Physiology & behavior 83: 361-371.

Dixson AF, Anderson MJ 2002. Sexual Selection, Seminal Coagulation and Copulatory Plug Formation in Primates. Folia Primatology 73: 6.

Dorus S, Evans PD, Wyckoff GJ, Choi SS, Lahn BT 2004. Rate of molecular evolution of the seminal protein gene SEMG2 correlates with levels of female promiscuity. Nature genetics 36: 1326-1329.

Doumas S, Kolokotronis A, Stefanopoulos P 2005. Anti-inflammatory and antimicrobial roles of secretory leukocyte protease inhibitor. Infection and immunity 73: 1271-1274.

Drannik AG, Henrick BM, Rosenthal KL 2011. War and peace between WAP and HIV: role of SLPI, trappin-2, elafin and ps20 in susceptibility to HIV infection. Biochemical Society transactions 39: 1427-1432.

Drapkin R, Horsten HHv, Lin Y, Mok SC, Crum CP, Welch WR, Hecht JL 2005. Human Epididymis Protein 4 (HE4) Is a Secreted Glycoprotein that Is Overexpressed by Serous and Endometrioid Ovarian Carcinomas. Cancer Res 65: 7.

Edstrom AM, Malm J, Frohm B, Martellini JA, Giwercman A, Morgelin M, Cole AM, Sorensen OE 2008. The major bactericidal activity of human seminal plasma is zinc-dependent and derived from fragmentation of the semenogelins. J Immunol 181: 3413-3421.

Excoffier L 2002. Human demographic history: refining the recent African origin model. Current opinion in genetics & development 12: 8.

Fay JC, Wu C-i 1999. A Human Population Bottleneck Can Account for the Discordance Between Patterns of Mitochondrial Versus Nuclear DNA Variation. Molecular Biology and Evolution 16: 1003-1005.

Ferrer-Admetlla A, Sikora M, Laayouni H, Esteve A, Roubinet F, Blancher A, Calafell F, Bertranpetit J, Casals F 2009. A natural history of FUT2 polymorphism in humans. Mol Biol Evol 26: 1993-2003.

Fledel-Alon A, Leffler EM, Guan Y, Stephens M, Coop G, Przeworski M 2011. Variation in human recombination rates and its genetic determinants. PLoS One 6: e20321.

Fledel-Alon A, Wilson DJ, Broman K, Wen X, Ober C, Coop G, Przeworski M 2009. Broad-scale recombination patterns underlying proper disjunction in humans. PLoS Genet 5: e1000658.

Frazer KA, Ballinger DG, Cox DR*, et al.* 2007. A second generation human haplotype map of over 3.1 million SNPs. Nature 449: 851-861.

Fu W, O'Connor TD, Jun G*, et al.* 2013. Analysis of 6,515 exomes reveals the recent origin of most human protein-coding variants. Nature 493: 216-220.

Fu Y-x 1997. Statistical Tests of Neutrality of Mutations Against Population Growth, Hitchhiking and Background Selection. Genetics 147: 10.

Fumagalli M, Cagliani R, Pozzoli U, Riva S, Comi GP, Menozzi G, Bresolin N, Sironi M 2009. Widespread balancing selection and pathogen-driven selection at blood group antigen genes. Genome research 19: 199-212.

Fumagalli M, Sironi M, Pozzoli U, Ferrer-Admetlla A, Pattini L, Nielsen R 2011. Signatures of environmental genetic adaptation pinpoint pathogens as the main selective pressure through human evolution. PLoS Genetics 7: e1002355.

Garamszegi LZ, Nunn CL 2011. Parasite-mediated evolution of the functional part of the MHC in primates. Journal of evolutionary biology 24: 184-195.

Ghosh M, Shen Z, Fahey JV, Cu-Uvin S, Mayer K, Wira CR 2010. Trappin-2/Elafin: a novel innate anti-human immunodeficiency virus-1 molecule of the human female reproductive tract. Immunology 129: 207-219.

Gonder MK, Disotell TR, Oates JF 2006. New Genetic Evidence on the Evolution of Chimpanzee Populations and Implications for Taxonomy. International Journal of Primatology 27: 1103-1127.

Gonder MK, Locatelli S, Ghobrial L, Mitchell MW, Kujawski JT, Lankester FJ, Stewart C-B, Tishkoff SA 2011. Evidence from Cameroon reveals differences in the genetic structure and histories of chimpanzee populations. PNAS 108: 12.

Gravel S, Henn BM, Gutenkunst RN, Indap AR, Marth GT, Clark AG, Yu F, Gibbs RA, Bustamante CD 2011. Demographic history and rare allele sharing among human populations. Proceedings of the National Academy of Sciences of the United States of America 108: 11983-11988.

Grossman SR, Andersen KG, Shlyakhter I*, et al.* 2013. Identifying recent adaptations in large-scale genomic data. Cell 152: 703-713.

Gutenkunst RN, Hernandez RD, Williamson SH, Bustamante CD 2009. Inferring the joint demographic history of multiple populations from multidimensional SNP frequency data. PLoS Genetics 5: e1000695.

Hai Yang Hu1. SG, Jiang Xi1, Zheng Yan1, Ning Fu2, Xiaoyu Zhang3, Corinna Menzel4, Hongyu, Liang3 HY, Min Zhao3, Rong Zeng2*, Wei Chen4,5, Svante Pa¨a¨bo6, Philipp Khaitovich1,6 2011. MicroRNA Expression and Regulation in Human, Chimpanzee, and Macaque Brains. PLoS Genet 7.

Hancock AM, Clark VJ, Qian Y, Di Rienzo A 2011. Population genetic analysis of the uncoupling proteins supports a role for UCP3 in human cold resistance. Molecular Biology and Evolution 28: 601-614.

Hancock AM, Witonsky DB, Gordon AS, Eshel G, Pritchard JK, Coop G, Di Rienzo A 2008. Adaptations to climate in candidate genes for common metabolic disorders. PLoS Genetics 4: e32.

Hartl DL, Clark AG. 2007. Principles of population genetics, 4th edition. Sunderland, Mass: Sinauer Associates.

Hedrick PW 2011. Population genetics of malaria resistance in humans. Heredity (Edinb) 107: 283-304.

Hellström I, Raycraft J, Hayden-Ledbetter M, Ledbetter JA, Schummer Ml, McIntosh M, Drescher C, Urban N, Hellström KE 2003. The HE4 (WFDC2) Protein Is a Biomarker for Ovarian Carcinoma. Cancer Research 63: 5.

Hennighausen LG, Sippel AE, Hobbs AA, Rosen JM 1982. Comparative sequence analysis of the mRNAs coding for mouse and rat whey protein. Nucleic Acids Res 10: 3733-3744.

Hernandez RD, Kelley JL, Elyashiv E, Melton SC, Auton A, McVean G, Sella G, Przeworski M 2011. Classic selective sweeps were rare in recent human evolution. Science 331: 920-924.

Hobolth A, Dutheil JY, Hawks J, Schierup MH, Mailund T 2011. Incomplete lineage sorting patterns among human, chimpanzee, and orangutan suggest recent orangutan speciation and widespread selection. Genome research 21: 349-356.

Holmes EC 2004. Adaptation and immunity. PLoS Biol 2: E307.

Hudson RR, Kreitman M, Aguadé M 1987. A Test of Neutral Molecular Evolution Based on Nucleotide Data. Genetics 116: 6.

Huff CD, Witherspoon DJ, Zhang Y*, et al.* 2012. Crohn's disease and genetic hitchhiking at IBD5. Mol Biol Evol 29: 101-111.

Hughes AL, Yeager M 1998. Natural selection at major histocompatibility complex loci of vertebrates. Annu Rev Genet 32: 415-435.

Hurle B, Swanson W, Program NCS, Green ED 2007. Comparative sequence analyses reveal rapid and divergent evolutionary changes of the WFDC locus in the primate lineage. Genome research 17: 276-286.

International HapMap Consortium 2005. A haplotype map of the human genome. Nature 437: 1299-1320.

Izumi K, Li Y, Zheng Y, Gordetsky J, Yao JL, Miyamoto H 2012. Seminal plasma proteins in prostatic carcinoma: increased nuclear semenogelin I expression is a predictor of biochemical recurrence after radical prostatectomy. Hum Pathol 43: 1991-2000.

Izumi K, Zheng Y, Miyamoto H 2011. Eppin expression in prostate cancer. Eur Urol 59: 1071-1072.

Jensen-Seaman MI, Li WH 2003. Evolution of the hominoid semenogelin genes, the major proteins of ejaculated semen. Journal of molecular evolution 57: 261-270.

Jobling M, Hurles M, Tyler-Smith C. 2004. Human evolutionary genetics : origins, peoples & disease: Garland Science.

Jonsson M, Frohm B, Malm J 2010. Binding of semenogelin I to intact human spermatozoa studied by flow cytometry and surface plasmon resonance. J Androl 31: 560-565.

Jonsson M, Lundwall A, Linse S, Frohm B, Malm J 2006. Truncated semenogelin I binds zinc and is cleaved by prostate-specific antigen. Journal of andrology 27: 542-547.

Kagedan D, Lecker I, Batruch I, Smith C, Kaploun I, Lo K, Grober E, Diamandis EP, Jarvi KA 2012. Characterization of the seminal plasma proteome in men with prostatitis by mass spectrometry. Clin Proteomics 9: 2.

Keinan A, Clark AG 2012. Recent explosive human population growth has resulted in an excess of rare genetic variants. Science 336: 740-743.

Kelley JL, Swanson WJ 2008. Positive selection in the human genome: from genome scans to biological significance. Annu Rev Genomics Hum Genet 9: 143-160.

Kim HL, Igawa T, Kawashima A, Satta Y, Takahata N 2010. Divergence, demography and gene loss along the human lineage. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365: 2451-2457.

King AE, Critchley HO, Kelly RW 2003. Innate immune defences in the human endometrium. Reprod Biol Endocrinol 1: 116.

Kingan SB, Tatar M, Rand DM 2003. Reduced polymorphism in the chimpanzee semen coagulating protein, semenogelin I. Journal of molecular evolution 57: 159-169.

Klein J, Sato A, Nagl S, O'hUigin C 1998. Molecular Trans-Species Polymorphism. Annual Review of Ecology and Systematics 29: 1-21.

Kosiol C, Vinar T, da Fonseca RR, Hubisz MJ, Bustamante CD, Nielsen R, Siepel A 2008. Patterns of positive selection in six Mammalian genomes. PLoS Genet 4: e1000144.

Kudaravalli S, Veyrieras JB, Stranger BE, Dermitzakis ET, Pritchard JK 2009. Gene expression levels are a target of recent natural selection in the human genome. Molecular Biology and Evolution 26: 649-658.

Laval G, Patin E, Barreiro LB, Quintana-Murci L 2010. Formulating a Historical and Demographic Model of Recent Human Evolution Based on Resequencing Data from Noncoding Regions. PLoS One 5: e10284.

Leffler EM, Gao Z, Pfeifer S*, et al.* 2013. Multiple Instances of Ancient Balancing Selection Shared Between Humans and Chimpanzees. Science.

Lilja H, Abrahamsson A, Lundwall A 1989. Semenogelin , the Predominant Protein in Human Semen. Biochemistry 264: 1894-1900.

Louicharoen C, Patin E, Paul R*, et al.* 2009. Positively selected G6PD-Mahidol mutation reduces Plasmodium vivax density in Southeast Asians. Science 326: 1546-1549.

Lundwall A 2007. A locus on chromosome 20 encompassing genes that are highly expressed in the epididymis. Asian journal of andrology 9: 540-544.

Lundwall A, Bjartell A, Olsson AY, Malm J 2002. Semenogelin I and II, the predominant human seminal plasma proteins, are also expressed in non-genital tissues. Mol Hum Reprod 8: 805-810.

Lundwall A, Clauss A 2011. Genes encoding WFDC- and Kunitz-type protease inhibitor domains: are they related? Biochemical Society transactions 39: 1398-1402.

Lundwall A, Lazure C 1995. A novel gene family encoding proteins with highly differing structure because of a rapidly evolving exon. FEBS Lett 374: 53-56.

Lundwall A, Ulvsback M 1996. The gene of the protease inhibitor SKALP/elafin is a member of the REST gene family. Biochem Biophys Res Commun 221: 323-327.

Ma G, Greenwell-Wild T, Lei K, Jin W, Swisher J, Hardegen N, Wild CT, Wahl SM 2004. Secretory leukocyte protease inhibitor binds to annexin II, a cofactor for macrophage HIV-1 infection. J Exp Med 200: 1337-1346.

Mackinnon MJ, Marsh K 2010. The Selection Landscape of Malaria Parasites. Science 328: 5.

Malm J, Jonsson M, Frohm B, Linse S 2007. Structural properties of semenogelin I. The FEBS journal 274: 4503-4510.

Manry J, Laval G, Patin E*, et al.* 2011. Evolutionary genetic dissection of human interferons. J Exp Med 208: 2747-2759.

Martellini JA, Cole AL, Venkataraman N, Quinn GA, Svoboda P, Gangrade BK, Pohl J, Sorensen OE, Cole AM 2009. Cationic polypeptides contribute to the anti-HIV-1 activity of human seminal plasma. FASEB J 23: 3609-3618.

McCrudden MT, Dafforn TR, Houston DF, Turkington PT, Timson DJ 2008. Functional domains of the human epididymal protease inhibitor, eppin. FEBS J 275: 1742-1750.

McDonald JH, Kreitman M 1991. Adaptative protein evolution at the Adh locus in Drosophila. Nature 351: 3.

McEvoy B, Beleza S, Shriver MD 2006. The genetic architecture of normal variation in human pigmentation: an evolutionary perspective and model. Human molecular genetics 15 Spec No 2: R176-181.

McKiernan PJ, McElvaney NG, Greene CM 2011. SLPI and inflammatory lung disease in females. Biochemical Society transactions 39: 1421-1426.

McNeely TB, Dealy M, Dripps DJ, Orenstein JM, Eisenberg SP, Wahl SM 1995. Secretory Leukocyte Protease Inhibitor: A Human Saliva Protein Exhibiting Anti-Human Immunodeficiency Virus 1 Activity In Vitro. The Journal of Clinical Investigation 96: 9.

McNeely TB, Shugars DC, Rosendahl M, Tucker C, Eisenberg SP, Wahl SM 1997. Inhibition of human immunodeficiency virus type 1 infectivity by secretory leukocyte protease inhibitor occurs prior to viral reverse transcription. Blood 90: 1141-1149.

McVean G, Spencer CC, Chaix R 2005. Perspectives on human genetic variation from the HapMap Project. PLoS Genet 1: e54.

Meyer WK, Arbeithuber B, Ober C, Ebner T, Tiemann-Boege I, Hudson RR, Przeworski M 2012. Evaluating the evidence for transmission distortion in human pedigrees. Genetics 191: 215-232.

Mills RE, Walter K, Stewart C*, et al.* 2011. Mapping copy number variation by population-scale genome sequencing. Nature 470: 59-65.

Moleirinho A, Seixas S, Lopes AM, Bento C, Prata MJ, Amorim A 2013. Evolutionary constraints in the beta-globin cluster: the signature of purifying selection at the delta-globin (HBD) locus and its role in developmental gene regulation. Genome Biol Evol 5: 559-571.

Moreau T, Baranger K, Dade S, Dallet-Choisy S, Guyot N, Zani ML 2008. Multifaceted roles of human elafin and secretory leukocyte proteinase inhibitor (SLPI), two serine protease inhibitors of the chelonianin family. Biochimie 90: 284-295.

Mueller JL, Ravi Ram K, McGraw LA, Bloch Qazi MC, Siggia ED, Clark AG, Aquadro CF, Wolfner MF 2005. Cross-species comparison of Drosophila male accessory gland protein genes. Genetics 171: 131-143.

Neafsey DE, Haas BJ 2011. 'Next-generation' sequencing becomes 'now-generation'. Genome Biology 12: 3.

Nielsen R 2005. Molecular Signatures of Natural Selection. Annu. Rev. Genet.

2005. 39:197–218 39: 24.

Nielsen R, Bustamante C, Clark AG*, et al.* 2005. A Scan for Positively Selected Genes in the Genomes of Humans and Chimpanzees. PLoS Biol 3.

Nielsen R, Hellmann I, Hubisz M, Bustamante C, Clark AG 2007. Recent and ongoing selection in the human genome. Nat Rev Genet 8: 857-868.

Nielsen R, Hubisz MJ, Hellmann I*, et al.* 2009. Darwinian and demographic forces affecting human protein coding genes. Genome research 19: 838-849.

Novembre J, Di Rienzo A 2009. Spatial patterns of variation due to natural selection in humans. Nature reviews. Genetics 10: 745-755.

Nunn CL 2002a. A comparative study of leukocyte counts and disease risk in primates. Evolution 56: 177–190.

Nunn CL 2002b. Spleen size, disease risk and sexual selection: a comparative study in primates. Evolutionary Ecology Research 4: 91–107.

Nunn CL, Altizer S, Jones KE, Sechrest W 2003. Comparative Tests of Parasite Species Richness in Primates. the american naturalist 163: 597–614.

Nunn CL, Gittleman JL, Antonovics J 2000. Promiscuity and the primate immune system. Science 290: 1168-1170.

O'Bleness M, Searles VB, Varki A, Gagneux P, Sikela JM 2012. Evolution of genetic and genomic features unique to the human lineage. Nature Reviews Genetics 13: 853-866.

O'Rand MG, Widgren EE 2012. Loss of Calcium in Human Spermatozoa via EPPIN, the Semenogelin Receptor. Biol Reprod 86: 55.

O'Rand MG, Widgren EE, Beyler S, Richardson RT 2009. Inhibition of human sperm motility by contraceptive anti-eppin antibodies from infertile male monkeys: effect on cyclic adenosine monophosphate. Biol Reprod 80: 279-285.

O'Rand MG, Widgren EE, Hamil KG, Silva EJ, Richardson RT 2011. Functional studies of eppin. Biochem Soc Trans 39: 1447-1449.

O'Rand MG, Widgren EE, Sivashanmugam P*, et al.* 2004. Reversible immunocontraception in male monkeys immunized with eppin. Science 306: 1189-1190.

O'Rand MG, Widgren EE, Wang Z, Richardson RT 2006. Eppin: an effective target for male contraception. Molecular and cellular endocrinology 250: 157-162.

O'Rand MG, Widgren EE, Hamil KG, Silva EJ, Richardson RT 2011. Epididymal Protein Targets: A Brief History of the Development of EPPIN as a Contraceptive. Journal of andrology.

Orr HA 2009. Fitness and its role in evolutionary genetics. Nature Reviews Genetics 10: 531-539.

Paasch U, Heidenreich F, Pursche T*, et al.* 2011. Identification of increased amounts of eppin protein complex components in sperm cells of diabetic and obese individuals by difference gel electrophoresis. Mol Cell Proteomics 10: M110 007187.

Patterson N, Richter DJ, Gnerre S, Lander ES, Reich D 2006. Genetic evidence for complex speciation of humans and chimpanzees. Nature 441: 1103-1108.

Peter A, Lilja H, Lundwall A, Malm J 1998. Semenogelin I and semenogelin II, the major gel-forming proteins in human semen, are substrates for transglutaminase. Eur. J. Biochem. 252: 5.

Petersen LC, Bjorn SE, Olsen OH, Nordfang O, Norris F, Norris K 1996. Inhibitory properties of separate recombinant Kunitz-type-protease-inhibitor domains from tissue-factor-pathway inhibitor. Eur J Biochem 235: 310-316.

Pritchard JK, Di Rienzo A 2010. Adaptation - not by sweeps alone. Nature reviews. Genetics 11: 665-667.

Pritchard JK, Pickrell JK, Coop G 2010. The genetics of human adaptation: hard sweeps, soft sweeps, and polygenic adaptation. Current biology : CB 20: R208-215.

Prugnolle F, Manica A, Charpentier M, Guegan JF, Guernier V, Balloux F 2005. Pathogen-driven selection and worldwide HLA class I diversity. Curr Biol 15: 1022-1027.

Przeworski M, Coop G, Wall JD 2005. The signature of positive selection on standing genetic variation. Evolution 59: 2312-2323.

Ptak SE, Hinds DA, Koehler K, Nickel B, Patil N, Ballinger DG, Przeworski M, Frazer KA, Paabo S 2005. Fine-scale recombination patterns differ between chimpanzees and humans. Nat Genet 37: 429-434.

Ptak SE, Roeder AD, Stephens M, Gilad Y, Paabo S, Przeworski M 2004. Absence of the TAP2 human recombination hotspot in chimpanzees. PLoS biology 2: e155.

Ptak SE, Voelpel K, Przeworski M 2004. Insights Into Recombination From Patterns of Linkage Disequilibrium in Humans. Genetics 167: 10.

Quach H, Barreiro LB, Laval G*, et al.* 2009. Signatures of purifying and local positive selection in human miRNAs. Am J Hum Genet 84: 316-327.

Raj T, Shulman JM, Keenan BT, Chibnik LB, Evans DA, Bennett DA, Stranger BE, De Jager PL 2012. Alzheimer disease susceptibility loci: evidence for a protein network under natural selection. American journal of human genetics 90: 720-726.

Rajesh A, Madhubabu G, Yenugu S 2011. Identification and characterization of Wfdc gene expression in the male reproductive tract of the rat. Molecular reproduction and development 78: 633-641.

Ramm SA, Oliver PL, Ponting CP, Stockley P, Emes RD 2008. Sexual selection and the adaptive evolution of mammalian ejaculate proteins. Molecular Biology and Evolution 25: 207-219.

Rayner JC, Liu W, Peeters M, Sharp PM, Hahn BH 2011. A plethora of Plasmodium species in wild apes: a source of human infection? Trends in parasitology 27: 222-229.

Richardson RT, Sivashanmugam P, Hall SH, Hamil KG, Moore PA, Ruben SM, French FS, O'Rand M 2001. Cloning and sequencing of human Eppin: a novel family of protease inhibitors expressed in the epididymis and testis. Gene 270: 93-102.

Roan NR, Muller JA, Liu H*, et al.* 2011. Peptides released by physiological cleavage of semen coagulum proteins form amyloids that enhance HIV infection. Cell host & microbe 10: 541-550.

Robert M, Gagnon C 1999. Semenogelin I: a coagulum forming, multifunctional seminal vesicle protein. Cell. Mol. Life Sci. 55: 16.

Roberts L 2011. 9 Billion? Science 333: 540-543.

Roghanian A, Sallenave JM 2008. Neutrophil elastase (NE) and NE inhibitors: canonical and noncanonical functions in lung chronic inflammatory diseases (cystic fibrosis and chronic obstructive pulmonary disease). J Aerosol Med Pulm Drug Deliv 21: 125-144.

Ronald J, Akey JM 2005. Genome-wide scans for loci under selection in humans. Human genomics 2: 113-125.

Sabeti PC, Reich DE, Higgins JM*, et al.* 2002. Detecting recent positive selection in the human genome from haplotype structure. Nature 419: 832-837.

Sabeti PC, Schaffner SF, Fry B*, et al.* 2006. Positive natural selection in the human lineage. Science 312: 1614-1620.

Sabeti PC, Varilly P, Fry B*, et al.* 2007. Genome-wide detection and characterization of positive selection in human populations. Nature 449: 913-918.

Sabeti PC, Walsh E, Schaffner SF*, et al.* 2005. The Case for Selection at CCR5-D32. 3.

Sallenave JM 2010. Secretory leukocyte protease inhibitor and elafin/trappin-2: versatile mucosal antimicrobials and regulators of immunity. Am J Respir Cell Mol Biol 42: 635-643.

Sallenave JM, Shulmann J, Crossley J, Jordana M, Gauldie J 1994. Regulation of secretory leukocyte proteinase inhibitor (SLPI) and elastase-specific inhibitor (ESI/elafin) in human airway epithelial cells by cytokines and neutrophilic enzymes. Am J Respir Cell Mol Biol 11: 733-741.

Scally A, Dutheil JY, Hillier LW*, et al.* 2012. Insights into hominid evolution from the gorilla genome sequence. Nature 483: 169-175.

Schaffner SF, Foo C, Gabriel S, Reich D, Daly MJ, Altshuler D 2005. Calibrating a coalescent simulation of human genome sequence variation. Genome research 15: 1576-1583.

Schalkwijk J, Wiedow O, Hirose S 1999. The trappin gene family: proteins defined by an N-terminal transglutaminase substrate domain and a C-terminal four-disulphide core. Biochem J 340 (Pt 3): 569-577.

Seger J, Smith WA, Perry JJ, Hunn J, Kaliszewska ZA, Sala LL, Pozzi L, Rowntree VJ, Adler FR 2010. Gene genealogies strongly distorted by weakly interfering mutations in constant environments. Genetics 184: 529-545.

Segurel L, Leffler EM, Przeworski M 2011. The case of the fickle fingers: how the PRDM9 zinc finger protein specifies meiotic recombination hotspots in humans. PLoS Biol 9: e1001211.

Seixas S, Ivanova N, Ferreira Z, Rocha J, Victor BL 2011. Loss and Gain of Function in SERPINB11: An Example of a Gene under Selection on Standing Variation, with Implications for Host-Pathogen Interactions. PLoS One 7.

Siepel A, Bejerano G, Pedersen JS*, et al.* 2005. Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res 15: 1034-1050.

Silva EJ, Hamil KG, Richardson RT, O'Rand MG 2012. Characterization of EPPIN's semenogelin I binding site: a contraceptive drug target. Biol Reprod 87: 56.

Smith VJ 2011. Phylogeny of whey acidic protein (WAP) four-disulfide core proteins and their role in lower vertebrates and invertebrates. Biochemical Society transactions 39: 1403-1408.

Stone AC, Battistuzzi FU, Kubatko LS, Perry GH, Jr., Trudeau E, Lin H, Kumar S 2010. More reliable estimates of divergence times in Pan using complete mtDNA sequences and accounting for population structure. Philosophical transactions of the Royal Society of London. Series B, Biological sciences 365: 3277-3288.

Suk EK, McEwen GK, Duitama J*, et al.* 2011. A comprehensively molecular haplotype-resolved genome of a European individual. Genome research 21: 1672-1685.

Swanson WJ 2003a. Adaptive evolution of genes and gene families. Curr Opin Genet Dev 13: 617-622.

Swanson WJ 2003b. Sex peptide and the sperm effect in Drosophila melanogaster. Proc Natl Acad Sci U S A 100: 9643-9644.

Swanson WJ, Clark AG, Waldrip-Dail HM, Wolfner MF, Aquadro CF 2001. Evolutionary EST analysis identifies rapidly evolving male reproductive proteins in Drosophila. Proc Natl Acad Sci U S A 98: 7375-7379.

Swanson WJ, Vacquier VD 2002. The rapid evolution of reproductive proteins. Nat Rev Genet 3: 137-144.

Swanson WJ, Wong A, Wolfner MF, Aquadro CF 2004. Evolutionary expressed sequence tag analysis of Drosophila female reproductive tracts identifies genes subjected to positive selection. Genetics 168: 1457-1465.

Tajima F 1989. Statistical method for testing the neutral mutation hypothesis by DNA polymorphism. Genetics 123: 585-595.

Takahata N 1993. Allelic genealogy and human evolution. Mol Biol Evol 10: 2-22.

Tennessen JA, Bigham AW, O'Connor TD*, et al.* 2012. Evolution and Functional Impact of Rare Coding Variation from Deep Sequencing of Human Exomes. Science 337: 64-69.

Tennessen JA, Madeoy J, Akey JM 2010. Signatures of positive selection apparent in a small sample of human exomes. Genome research 20: 1327-1334.

Terai K, Yoshida K, Yoshiike M, Fujime M, Iwamoto T 2010. Association of seminal plasma motility inhibitors/semenogelins with sperm in asthenozoospermia-infertile men. Urologia internationalis 85: 209-215.

Thacker S, Yadav SP, Sharma RK, Kashou A, Willard B, Zhang D, Agarwal A 2011. Evaluation of sperm proteins in infertile men: a proteomic approach. Fertility and sterility 95: 2745-2748.

The 1000 Genomes Project Consortium 2010. A map of human genome variation from population-scale sequencing. Nature 467: 1061-1073.

Thompson EE, Kuttab-Boulos H, Witonsky D, Yang L, Roe BA, Di Rienzo A 2004. CYP3A variation and the evolution of salt-sensitivity variants. Am J Hum Genet 75: 1059-1069.

Thorisson GA, Smith AV, Krishnan L, Stein LD 2005. The International HapMap Project Web site. Genome Res 15: 1592-1593.

Tillier ER, Charlebois RL 2009. The human protein coevolution network. Genome research 19: 1861-1871.

Tishkoff SA, Reed FA, Ranciaro A*, et al.* 2007. Convergent adaptation of human lactase persistence in Africa and Europe. Nat Genet 39: 31-40.

Tsunemi M, Matsuura Y, Sakakibara S, Katsube Y 1996. Crystal structure of an elastase-specific inhibitor elafin complexed with porcine pancreatic elastase determined at 1.9 A resolution. Biochemistry 35: 11570-11576.

Vernot B, Stergachis AB, Maurano MT, Vierstra J, Neph S, Thurman RE, Stamatoyannopoulos JA, Akey JM 2012. Personal and population genomics of human regulatory variation. Genome Res 22: 1689-1697.

Verrelli BC, Lewis CM, Jr., Stone AC, Perry GH 2008. Different selective pressures shape the molecular evolution of color vision in chimpanzee and human populations. Molecular Biology and Evolution 25: 2735-2743.

Verrelli BC, Tishkoff SA, Stone AC, Touchman JW 2006. Contrasting histories of G6PD molecular evolution and malarial resistance in humans and chimpanzees. Molecular Biology and Evolution 23: 1592-1601.

Voight BF, Adams AM, Frisse LA, Qian Y, Hudson RR, Di Rienzo A 2005. Interrogating multiple aspects of variation in a full resequencing data set to infer human population size changes. Proc Natl Acad Sci U S A 102: 18508-18513.

Voight BF, Kudaravalli S, Wen X, Pritchard JK 2006. A map of recent positive selection in the human genome. PLoS biology 4: e72.

Wahl SM, McNeely TB, Janoff EN, Shugars D, Worley P, Tucker C, Orenstein JM 1997. Secretory leukocyte protease inhibitor (SLPI) in mucosal fluids inhibits HIV-I. Oral Dis 3 Suppl 1: S64-69.

Wang Z, Widgren EE, Richardson RT, O'Rand MG 2007a. Characterization of an eppin protein complex from human semen and spermatozoa. Biol Reprod 77: 476-484.

Wang Z, Widgren EE, Richardson RT, Orand MG 2007b. Eppin: a molecular strategy for male contraception. Soc Reprod Fertil Suppl 65: 535-542.

Wang Z, Widgren EE, Sivashanmugam P, O'Rand MG, Richardson RT 2005. Association of eppin with semenogelin on human spermatozoa. Biology of reproduction 72: 1064-1070.

Ward LD, Kellis M 2012. Evidence of abundant purifying selection in humans for recently acquired regulatory functions. Science 337: 1675-1678.

Wegmann D, Excoffier L 2010. Bayesian inference of the demographic history of chimpanzees. Molecular Biology and Evolution 27: 1425-1435.

Weldon S, McGarry N, Taggar CC, McElvaney NG 2007. The role of secretory leucoprotease inhibitor in the resolution of inflammatory responses. Biochemical Society transactions 35: 3.

Weldon S, Taggart CC 2007. Innate host defense functions of secretory leucoprotease inhibitor. Experimental lung research 33: 485-491.

Wilkinson TS, Roghanian A, Simpson AJ, Sallenave JM 2011. WAP domain proteins as modulators of mucosal immunity. Biochemical Society transactions 39: 1409-1415.

Williams SE, Brown TI, Roghanian A, Sallenave JM 2006. SLPI and elafin: one glove, many fingers. Clinical science 110: 21-35.

Williamson S, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R 2005. Localizing recent adaptive evolution in the human genome. PLoS Genetics preprint: e90.

Williamson SH, Hubisz MJ, Clark AG, Payseur BA, Bustamante CD, Nielsen R 2007. Localizing recent adaptive evolution in the human genome. PLoS Genet 3: e90.

Wlasiuk G, Nachman MW 2010a. Adaptation and constraint at Toll-like receptors in primates. Molecular Biology and Evolution 27: 2172-2186.

Wlasiuk G, Nachman MW 2010b. Promiscuity and the rate of molecular evolution at primate immunity genes. Evolution 64: 2204-2220.

Wolfner M 2002. The gifts that keep on giving: physiological functions and evolutionary dynamics of male seminal proteins in Drosophila. Heredity 88: 85-93.

Woolfe A, Goodson M, Goode DK*, et al.* 2004. Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development. PLoS Biol 3: e7.

Yang Z 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. Mol Biol Evol 24: 1586-1591.

Yenugu S, Richardson RT, Sivashanmugam P, Wang Z, O'Rand M G, French FS, Hall SH 2004. Antimicrobial activity of human EPPIN, an androgen-regulated, sperm-bound protein with a whey acidic protein motif. Biology of reproduction 71: 1484-1490.

Zeng K, Fu YX, Shi S, Wu CI 2006. Statistical tests for detecting positive selection by utilizing high-frequency variants. Genetics 174: 1431-1439.

Zhang J, Nielsen R, Yang Z 2005. Evaluation of an improved branch-site likelihood method for detecting positive selection at the molecular level. Molecular Biology and Evolution 22: 2472-2479.

Zhao H, Lee WH, Shen JH, Li H, Zhang Y 2008. Identification of novel semenogelin I-derived antimicrobial peptide from liquefied human seminal plasma. Peptides 29: 505-511.

Appendices

Supplementary Material Article 1

Differing evolutionary histories of *WFDC8* (short-term balancing) in Europeans and *SPINT4* (incomplete selective sweep) in Africans

Mol Biol Evol. 2011 Oct;28(10):2811-22

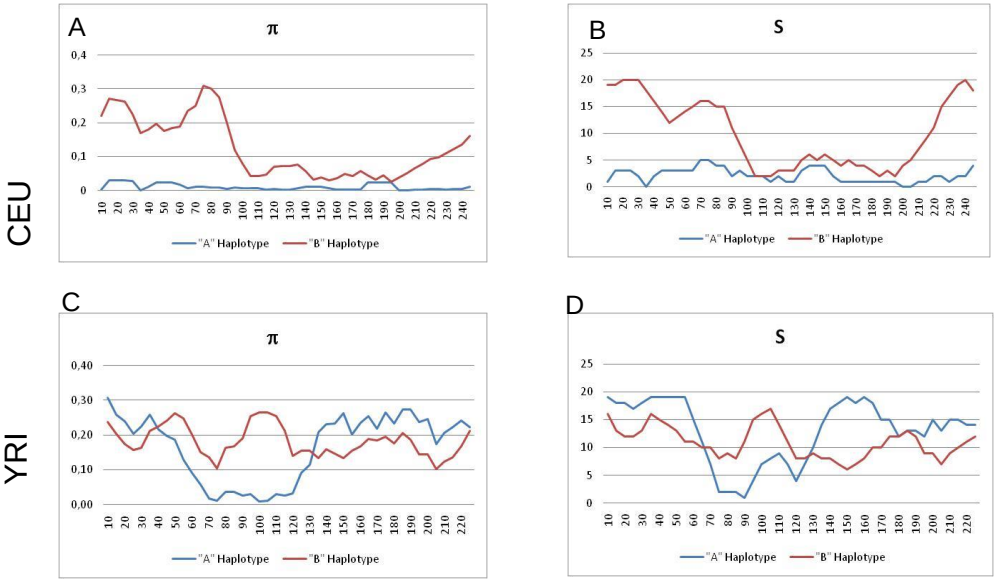
Supplementary Tables

Supplementary Table S1: HapMap Phase II individuals (Coriell repository) selected for the re-sequencing and genotyping studies. Samples that were typed for the SNP rs7273669 using the BseRI restriction enzyme assay (but not re-sequenced) are indicated by an asterisk (*).

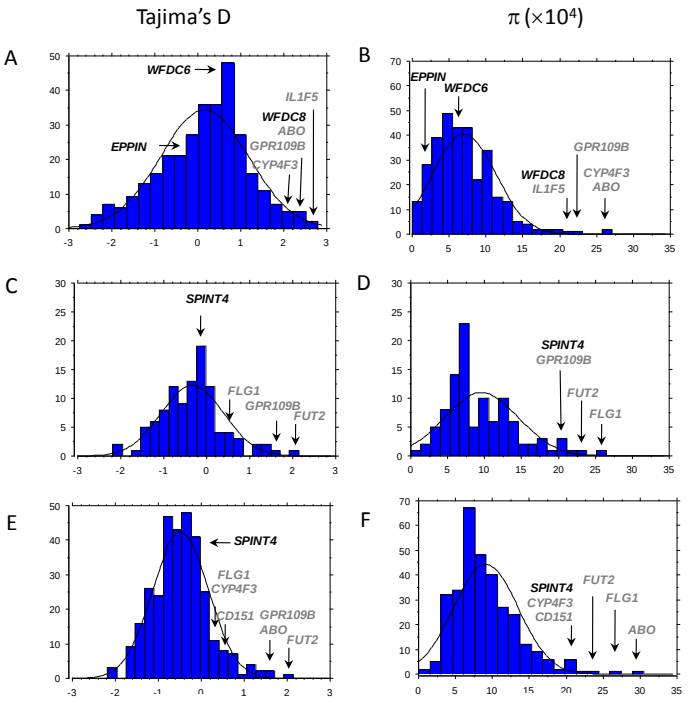
Population	Individuals
CEU	NA6985, NA12751, NA12875, NA12003, NA12006, NA12156, NA11994, NA12057, NA12248, NA11993, NA07056, NA06994, NA07022, NA07000, NA12155, NA12144, NA12004, NA12005, NA11881, NA11831, NA12056, NA12145, NA12249, NA06993,* NA07034,* NA07055,* NA07345,* NA07357,* NA11829,* NA11830,* 6a14,*16,* NA11832,* NA11839,* NA11840,* NA11882,* NA11992,* NA11995,* 13a27,*28,* NA12043,* NA12044,* NA12146,* NA12154,* NA12234,* NA12236,* NA12239,* NA12264,* NA12716,* NA12717,* NA12750,* NA12760,* NA12761,* NA12762,* NA12763,* NA12812,* NA12813,* NA12814,* NA12815,* NA12872,* NA12873,* NA12874,* NA18501,* NA18502,* NA18504,* NA18505,* NA18508,* NA18516,* NA18522,* NA18852,* NA18855,* NA18858,* NA18859,* NA18862,* NA18871,* NA18912,* NA19099,* NA19101,* NA19127,* NA19128,* NA19137,* NA19140,* NA19159,* NA19238,* NA12891,* NA12892*
Yoruba	NA_18501, NA_18502, NA_18504, NA_18505, NA_18508, NA_18516, NA_18522, NA_18852, NA_18855, NA_18858, NA_18859, NA_18862, NA_18871, NA_18912, NA_19099, NA_19101, NA_19127, NA_19128, NA_19137, NA_19140, NA_19159, NA_19238, NA18507*, NA18517*, NA18523*, NA18853*, NA18856*, NA18861*, NA18870*, NA18913*, NA19092*, NA19093*, NA19098*, NA19102*, NA19116*, NA19119*, NA19130*, NA19131*, NA19138*, NA19141*, NA19143*, NA19144*, NA19152*, NA19153*, NA19238*, NA19160*, NA19171*, NA19172*, NA19192*, NA19193*, NA19200*, NA19201*, NA19203*, NA19204*, NA19206*, NA19207*, NA19209*, NA19210*, NA19222*, NA19223*, NA19238*, NA19239
JPT+CHB	NA_18576*, NA_18943*, NA_19000*, NA_18978*, NA_18940*, NA_18635*, NA_18620*, NA_18572*, NA_18992*, NA_18563*, NA_18573*, NA_18545*, NA_18594*, CD18609*, NA_18577*, NA_18566*, NA_18948*, NA_18555*, NA_18529*, NA_18582*, NA_18532*, NA_18524*, NA_18998*, NA_18636*, NA_18633*, NA_18611*, NA_18997*, NA_18593*, NA_18552*, NA_18947*, NA_18999*, NA_18990*, NA_18980*, NA_18637*, NA_18561*, NA_18558*, NA_18537*, NA_19007*, NA_18991*, NA_18579*, NA_18967*, NA_18969*, NA_19012*, NA_19003*, NA_18987*, NA_18975*, NA_18956*, NA_18571*, NA_18976*, NA_19005*, NA_18961*, NA_18949*, NA_18562*, NA_18592*, NA_18951*, NA_18942*, NA_18995*, NA_18964*, NA_18959*, NA_18603*, NA_18953*, NA_18526*, NA_18547*, NA_18540*, NA_18994*, NA_18974*, NA_18550*, NA_18968*, NA_18944*, NA_18966*, NA_18972*, NA_18970*, NA_18960*, NA_18965*, NA_18632*, NA_18624*, NA_18612*, NA_18608*, NA_18605*, NA_18945*, NA_18981*, NA_18973*, NA_18952*, NA_18622*,

NA_18542*, NA_18623*, NA_18621*, NA_18570*, NA_18564*

Supplementary Figures

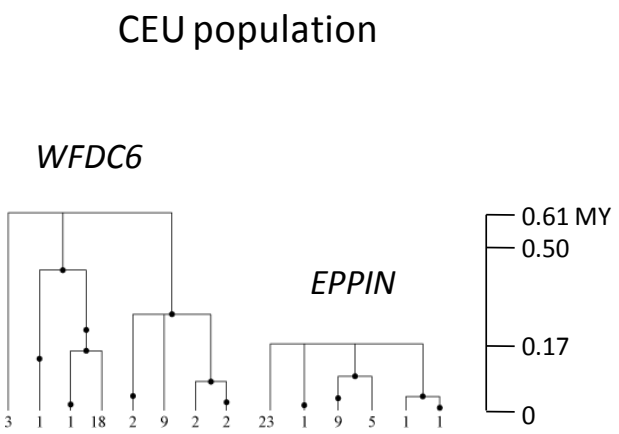


Supplementary Figure S1: Sliding windows of diversity for HapMap Phase II haplotypes. "A" and "B" correspond to the 200 kb cluster region centered on *EPPIN* for CEU and C and D to the 200 kb cluster region centered on *SPINT4* for YRI. The window length used was 20 SNPs with a step size of 5 SNPs. The window midpoint is indicated in x-axis. *WFDC8* extends from 145 to 224 SNPs and *SPINT4* from 63 to 71 SNPs. π - genetic diversity; S- number of variable SNPs.

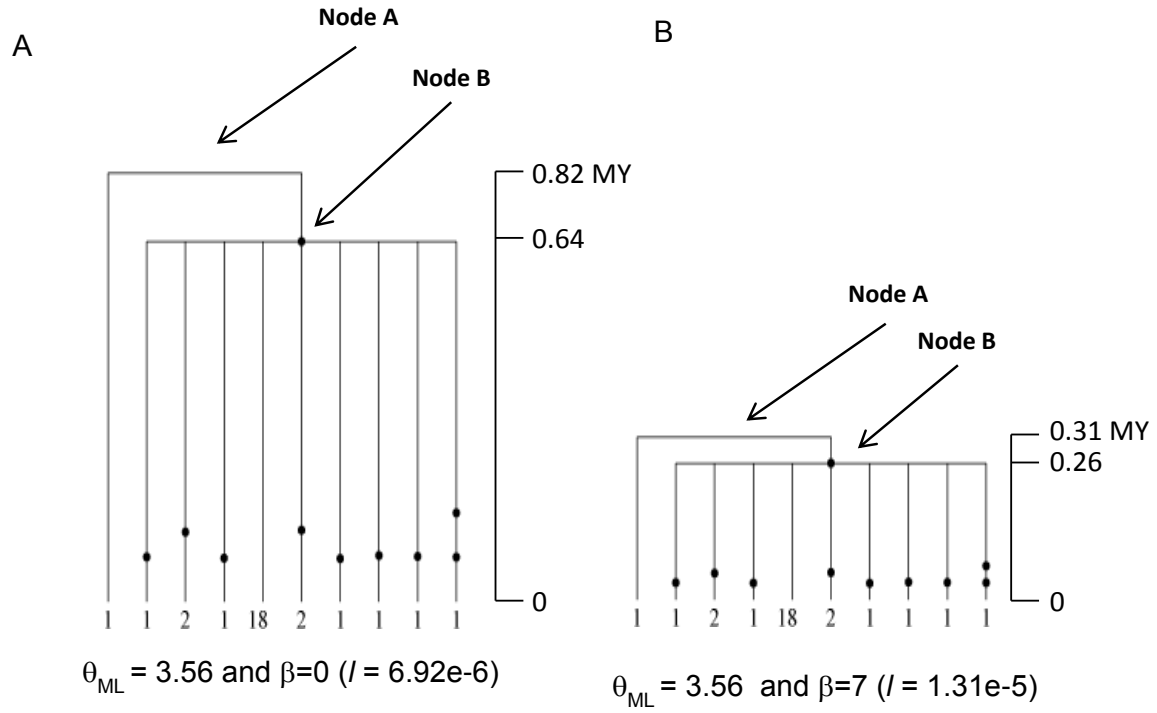


Supplementary Figure S2: Empirical distributions of Tajima's D and π built using the 316 genes surveyed by SeattleSNPs (http://pga.gs.washington.edu/. Genes within the upper extreme of the distribution are marked in grey. Gene classes with values close to survey genes are indicated by *WFDC* gene name. A and B: CEU samples –

SeattleSNPs panels 1 to 3. C and D: YRI sample – SeattleSNPs panel 2. E and F: African-American (AA) descent samples – SeattleSNPs panels 1 to 3.



Supplementary Figure S3: Gene genealogy tree of *WFDC6* and *EPPIN* as estimated by Genetree. Time is scaled in millions of years (MY).



Supplementary Figure S4: **Gene genealogy trees as estimated by Genetree for *SPINT4* branch corresponding to Ser73+98A haplotype. Tree calculated assuming no selection (A). Tree calculated assuming the maximum likelihood of the**

Appendix II

Supplementary Material Article 2

Reproduction and Immunity Driven Natural Selection in the Human WFDC Locus

Mol Biol Evol. 2013 Apr;30(4):938-50

Supplementary Tables

Supplementary Table S1: Regions of the genome sequenced.

Chr 20 coordinates*	Gene
	WFDC-CEN
43,171,507-43,177,217	*WFDC5* (WAP four-disulfide core domain 5)
43,185,481-43,186,520	*WFDC12* (WAP four-disulfide core domain 12)
43,236,912-43,238,599	*PI3* (elafin)
43,269,088-43,271,823	*SEMG1* (semenogelin 1)
43,269,088-43,286,513	*SEMG2* (semenogelin 2)
43,314,293-43,316,620	*SLPI* (secretory leukocyte peptidase inhibitor)
	WFDC-TEL
43,541,899-43,543,586	*WFDC2* (WAP four-disulfide core domain 2)
43,574,515-43,577,678	*SPINT3*
43,596,250-43,601,548	*WFDC6* (WAP four-disulfide core domain 6)
43,602,679-43,609,442	*SPINLW1* (WAP four-disulfide core domain 7)
43,613,815-43,641,379	*WFDC8* (WAP four-disulfide core domain 8)
43,669,992-43,693,321	*WFDC9* (WAP four-disulfide core domain 9)
43,691,799-43,693,245	*WFDC10a* (WAP four-disulfide core domain 10a)
43,710,616-43,732,292	*WFDC11* (WAP four-disulfide core domain 11)
43,746,704-43,767,072	*WFDC10b* (WAP four-disulfide core domain 10b)
43,764,069-43,770,870	*WFDC13* (WAP four-disulfide core domain 13)
43,784,402-43,787,749	*SPINT4*
43,836,254-43,853,954	*WFDC3* (WAP four-disulfide core domain 3)

[#]Control Regions (from Andres, 2010)

Ψ ID	Chrromosome coordinate	Gene and chr of origin			Processed Ψs per genome			
					Human	Chimp	Orang	Rhesus
ENCODE								
Ψ:79794	chr11:5505093-5505962	Unknown			1	1	1	1
NON-ENCODE								
Ψ:89	chr1:38020719-38021144	C14ORF138	@	Chr14	1	1	1	0
Ψ:127271	chr1:224666087-224666538	Unknown			1	3	2	4
Ψ:511	chr1:181191136-181191557	NCR1	@	Chr19	1	1	1	1
Ψ:716	chr1:245414326-245414818	PEX5	@	Chr12	1	1	1	1
Ψ:189	chr1:61892410-61893166	SPR	@	Chr2	1	1	1	1
Ψ:414	chr1:156412265-156414162	ELL2	@	Chr5	3	3	3	3
Ψ:725	chr2:7794021-7794677	PSMB1	@	Chr6	1	1	ORF	2?
Ψ:881	chr2:76338812-76339387	NP	@	Chr14	1	1	1	1 + ORF
Ψ:919	chr2:101503725-101504888	PRCP	@	Chr11	1	1	1	1
Ψ:10377	chr2:146951393-146952148	Unknown			1	1	1	1
Ψ:10800	chr2:158526147-158527646	MTA3	@	Chr2	1	1	1	ND

Ψ:1317	chr3:34889168-34890350	FECH	@	Chr18	1	1	1	1
Ψ:1390	chr3:59496630-59497181	Unknown			2	2	2	2
Ψ:131502	chr3:81498830-81499811	Unknown			1	1	1	ND
Ψ:1588	chr3:146885050-146885571	GM2A	@	Chr5	2	2	2	2
Ψ:75498	chr4:36184061-36186883	FBXO38	@	Chr5	1	1	1	1
Ψ:1886	chr4:89667737-89668341	CD53	@	Chr1	1	1	1	1
Ψ:18262	chr4:176732748-176734338	ADAM29	@	Chr4	2	2	2	2
Ψ:1835	chr4:68730627-68731081	Unknown			1	1	1	1
Ψ:2117	chr5:29141432-29142101	C14ORF45	@	Chr14	1	1	1	1
Ψ:19453	chr5:133084092-133084524	DPH4	@	Chr11	1	1	1	1
Ψ:2934	chr6:139701013-139702410	DNAJC7	@	Chr17	2	2	2	3
Ψ:3153	chr7:55279435-55280868	SLC19A3	@	Chr2	2	2	2	2
Ψ:20740	chr7:128120741-128121217	IMP3	@	Chr15	2	2	2	2
Ψ:3607	chr8:60632574-60632998	NUDT15	@	Chr13	2	2 + 1 ORF	2	2
Ψ:3693	chr8:97208019-97209206	Unknown			1	1	1	1
Ψ:69522	chr9:6185935-6186853	GTF3A	@	Chr13	5	5	5	5
Ψ:21854	chr9:21685170-21686842	KHSRP	@	Chr19	1	1	1	1
Ψ:4067	chr9:106130426-106131316	WDR45L	@	Chr17	3	2 + 1 ORF	2 + 1 ORF	2
Ψ:4320	chr10:79498396-79499304	GNAI2	@	Chr3	3	3	2	1
Ψ:4547	chr11:40061840-40062643	ZCCHC9	@	Chr5	3	2	3	3
Ψ:4837	chr11:113829007-113830057	CCRN4L	@	Chr4	1	1	1	2
Ψ:4765	chr11:91708853-91709310	NDUFB11	@	ChrX	1	1	1	1 + 1 ORF
Ψ:4913	chr12:7650698-7651196	CLTA	@	Chr9	2	2	2	2
Ψ:5244	chr12:107199014-107200142	APOBEC3F	@	Chr22	1	1	1	1
Ψ:5055	chr12:44124847-44125485	MESDC2	@	Chr15	1	1	1	1
Ψ:73106	chr13:24217272-24218374	IRX1	@	Chr5	1	1	1	1
Ψ:5504	chr13:89442087-89443147	PEX12	@	Chr17	1	1	1	0
Ψ:5477	chr13:74300713-74302220	RIOK3	@	Chr18	1	1	1	1
Ψ:5700	chr14:61368336-61368840	COX4I1	@	Chr16	2	1	2	2 + 1 ORF
Ψ:62081	chr14:88646206-88647320	MPPE1	@	Chr18	2	1	1	1
Ψ:6310	chr16:82710324-82711542	PLK-1	@	Chr16	2	2	2	2
Ψ:6549	chr17:66136577-66137001	Unknown			1	1	1	1
Ψ:26649	chr19:18012148-18012866	APOA1BP	@	Chr1	2	2	2	2
Ψ:6997	chr20:21429964-21430620	GSTM3	@	Chr1	1	1	1	2
Ψ:7045	chr20:37391075-37391973	ATG3	@	Chr3	1	2	1	1
Ψ:7112	chr21:15051781-15052541	POLR2C	@	Chr16	1	1	1	1
Ψ:7176	chr21:45316169-45316649	Unknown			1	1	1	1

* Chromosome coordinates are based in the human march 2006 assembly (hg18; genome.ucsc.edu)

[#] These loci were selected using the following filters: Processed ψs (according to the annotation in ψ.org); Minimum length of 400 bp; Unlinked to each other; ot OR, not ribosomal RNA (according to the annotation in ψ.org); With chimpanzee, orangutan and rhesus orthologues; No overlap with UCSC genes; No overlap with highly conserved regions (most conserved, UCSC); Average genome recombination rate; Processed ψs in single copy in human and

chimpanzee genomes were preferred. (When this was not possible, one ψ per family was accepted provided that the members of the family had 90% identity or less among them.)

Supplementary Table S1**: HapMap Phase I/II samples sequenced.**

Population	**Sample numbers**
YRI (25)	NA18502, NA18517, NA18507, NA18523, NA18505, NA18861, NA18508, NA18858, NA18516, NA18522, NA18501, NA19102, NA19172, NA19116, NA18504, NA18853, NA18856, NA18870, NA18871, NA18912, NA19093, NA19137, NA19138, NA19171, NA19200
CEU (21)	NA11832, NA11840, NA11993, NA12004, NA12156, NA12813, NA11995, NA12006, NA06985, NA12044, NA11829, NA11830, NA11831, NA11992, NA11994, NA12154, NA12155, NA12003, NA12005, NA11840, NA12813
CHB (20)	NA18526, NA18529, NA18532, NA18537, NA18540, NA18542, NA18545, NA18547, NA18550, NA18552, NA18555, NA18564, NA18566, NA18570, NA18571, NA18573, NA18576, NA18577, NA18579, NA18582
JPT (5)	NA18942, NA18947, NA18956, NA18980, NA18992

Supplementary Table S2: a) SNPs nonsynonymous b) Fixed Differences nonsynonymous analysis with SIFT and PolyPhen v2.

a) Nonsynonymous and Stop mutations

SNP ID	Protein	Frequency	Residue	Expected Codon	Found Codon	SIFT	PolyPhen
rs79417830	WFDC5	0.05	S103L	TCG	TTG	Tolerated	Benign
rs17422688	WFDC5	0.05	H97Y	CAC	TAC	Tolerated	Benign
rs115077112	WFDC12	0.01	A20V	GCT	GTT	Affect protein function	Possibly damaging
rs17333103	PI3	0.08	T17M	ACG	ATG	Tolerated	Benign
rs2664581	PI3	0.08	T34P	ACT	CCT	Tolerated	Benign
rs151106234	PI3	0.01	K36N	AAA	AAT	Tolerated	Benign
	SEMG1	0.01	G26D	GGT	GAT	Affect protein function	Possibly damaging
rs2301366	SEMG1	0.33	S79T	TCC	ACC	Tolerated	Benign
rs61729393	SEMG1	0.02	L100Q	CTA	CAA	Affect protein function	Probably damaging
rs75447646	SEMG1	0.01	S193N	AGT	AAT	Tolerated	Benign
rs141417035	SEMG1	0.01	R208C	CGT	TGT	Affect protein function	Probably damaging
rs2233886	SEMG1	0.02	S312R	AGC	CGC	Affect protein function	
rs139222986	SEMG1	0.02	Q327R	CAA	CGA	Tolerated	Benign
rs61730001	SEMG1	0.02	Q357R	CAA	CGA	Tolerated	Benign
rs151126678	SEMG1	0.02	H359Q	CAC	CAA	Tolerated	Benign
	SEMG1	0.01	R372C	CGC	TGC	Tolerated	Benign
rs2233887	SEMG1	0.02	R372L	CGC	CTC	Affect protein function	Benign
	SEMG1	0.01	R447H	CGT	CAT	Tolerated	Benign
	SEMG2	0.01	S34R	AGC	AGT	Tolerated	Benign
rs2233896	SEMG2	0.07	Q43K	CAA	AAA	Tolerated	Benign
rs2233897	SEMG2	0.04	T57A	ACT	GCT	Tolerated	Probably damaging
rs144655663	SEMG2	0.01	R208L	CGT	CTT	Tolerated	Benign
rs2233901	SEMG2	0.1	S274N	AGT	AAT	Tolerated	Benign
rs2233903	SEMG2	0.08	H279Y	CAT	TAT	Affect protein function	Probably damaging
	SEMG2	0.01	D336Y	GAT	TAT	Tolerated	Probably damaging
rs2071650	SEMG2	0.1	G368R	GGT	CGT	Tolerated	Possibly damaging
rs144708311	SEMG2	0.01	R417H	CGT	CAT	Tolerated	Possibly damaging
rs116547998	SEMG2	0.02	E463K	GAA	AAA	Tolerated	Probably damaging
rs148116830	SEMG2	0.01	E475*	GAA	TAA	Stop	
rs6032259	SPINT3	0.45	L77S	CGA	GAA	Tolerated	Benign
rs73305953	SPINT3	0.06	F54S	TTC	TCC	Affect protein function	Probably damaging
rs118187173	WFDC8	0.02	P155Q	CCA	CAA	Affect protein function	Benign
rs2250860	WFDC8	0.26	N137S	AAT	AGT	Tolerated	Benign

rs2272955	WFDC8	0.17	M96T	ATG	ACG	Tolerated	Benign
	WFDC8	0.01	K73N	AAG	AAT	Tolerated	Benign
rs79465015	WFDC8	0.02	M39T	ATG	ACG	Tolerated	Benign
	WFDC9	0.01	D29G	GAT	GGT	Tolerated	Benign
rs2245898	WFDC9	0.42	N27T	AAC	ACC	Tolerated	Benign
rs16990631	SPINT4	0.08	A30E	GCG	GAG	Tolerated	Possibly damaging
rs6017667	SPINT4	0.44	G73S	GGC	AGC	Affect protein function	Probably damaging
	SPINT4	0.01	P98S	CCA	TCA	Tolerated	Benign
rs58353703	WFDC10A	0.01	Q30L	CAG	CTG	Affect protein function	Possibly damaging
rs232729	WFDC10B	0.13	L8P	CTT	CCT	Tolerated	Benign
rs76083103	WFDC3	0.04	N217K	AAC	AAG	Tolerated	Benign
rs80158904	WFDC3	0.02	R81G	CGG	GGG	Tolerated	Benign
	WFDC3	0.48	D75N	GAT	AAT	Tolerated	Possibly damaging
	WFDC3	0.45	R74K	AGA	AAA	Tolerated	Benign
rs73122754	WFDC3	0.01	R63L	CGG	CTG	Tolerated	Benign
rs6032538	WFDC3	0.36	H36D	CAT	GAT	Tolerated	Benign

b) Fixed Differences nonsynonymous

Position (chr20)	Human	Chimp	Orang	Macaque	AA subs	Gene	PolyPhen	SIFT
43186462	V	A	A	A	V14A	WFDC12	Benign	Affect protein function
43186463	V	I	I	I	V14I	WFDC12	Benign	Affect protein function
43269459	F	S	S	S	F36 S	SEMG1	Benign	Tolerated
43269641	Q	E	E	E	Q97E	SEMG1	Benign	Tolerated
43270493	L	I	?	?	L381I	SEMG1	Benign	Tolerated
43270494	L	Q	?	?	L381Q	SEMG1	Benign	Tolerated
43270499	A	P	?	?	A383P	SEMG1	Benign	Tolerated
43270673	Q	E	?	?	Q441E	SEMG1	Benign	Tolerated
43283919	W	R	R	R	W78R	SEMG2	Benign	Tolerated
43284163	C	Y	Y	Y	C159Y	SEMG2	Benign	Tolerated
43284553	P	Q	Q	Q	P289Q	SEMG2	Benign	Tolerated
43284582	H	N	N	N	H299N	SEMG2	Benign	Tolerated
43285180	I	T	T	T	I494T	SEMG2	Benign	Tolerated
43285287	K	E	E	E	K534E	SEMG2	Benign	Tolerated
43600054	P	L	L	L	P62L	WFDC6	Benign	Tolerated
43624268	F	L	L	L	F11L	WFDC8	Benign	Tolerated
43672210	V	I	I	M	V9I	WFDC9	Benign	Tolerated
43710782	G	E	E	E	G85E	WFDC11	Benign	Tolerated
43746903	I	V	V	V	I84V	WFDC10B	Benign	Tolerated

Supplementary Table S3: Summary statistics for *WFDC* locus genes.

Gene	Pop	[a]S	Length	[b]π	[c]Θ$_w$	[d]*D*	*P*-value	[e] *D**	*P*-value	[f]*H'*	*P*-value	[g]p (HKA)	[h]p (MWU$_{high}$)
WFDC5	CHB+JPT	15	3787	1.85	3.35	-1.37	0.149	-2.11	0.12	0.17	0.496	0.351	0.945
WFDC12	CHB+JPT	9	2756	0.98	2.02	-1.45	0.099	-1.95	0.09	-0.43	0.316	0.882	0.928
PI3	CHB+JPT	17	4546	1.68	3.82	-1.76	0.041	-3.40	0.12	-3.79	0.018	0.692	0.966
SEMG1	CHB+JPT	9	3964	0.76	2.02	-1.75	0.058	-2.60	0.14	-1.79	0.084	0.928	0.928
SEMG2	CHB+JPT	11	3658	0.93	2.47	-1.82	0.048	-3.16	0.24	-3.55	0.018	0.429	0.978
SLPI	CHB+JPT	16	5366	2.76	3.60	-0.72	0.333	0.72	0.08	-2.85	0.037	0.841	0.680
WFDC2	CHB+JPT	6	3978	1.48	1.35	0.26	0.360	0.32	0.28	-0.42	0.313	0.931	0.333
SPINT3	CHB+JPT	4	3722	1.04	0.91	0.34	0.277	1.00	0.04	0.91	0.030	0.207	0.265
WFDC6	CHB+JPT	9	2803	1.62	2.02	-0.57	0.431	0.03	0.22	-0.24	0.345	0.371	0.462
EPPIN	CHB+JPT	12	3431	2.72	2.68	0.05	0.151	1.48	0.00	0.05	0.499	0.091	0.101
WFDC8	CHB+JPT	31	7169	5.01	7.01	-0.97	0.167	1.06	0.06	-3.71	0.019	0.126	0.852
WFDC9/10A	CHB+JPT	24	6855	8.11	5.39	1.65	0.079	1.11	0.05	-0.48	0.347	0.694	0.048
WFDC11	CHB+JPT	13	5032	4.23	2.94	1.34	0.139	1.51	0.01	0.54	0.205	0.236	0.087
WFDC10B/13	CHB+JPT	17	7356	4.42	3.81	0.50	0.175	-0.05	0.22	-0.80	0.232	0.560	0.159
SPINT4	CHB+JPT	14	3521	5.03	3.26	1.72	0.077	1.06	0.07	0.72	0.107	0.049	0.075
WFDC3	CHB+JPT	45	7562	13.50	10.10	1.15	0.160	0.62	0.12	0.46	0.228	0.030	0.131
WFDC5	CEU	8	3787	2.08	2.02	0.09	0.368	0.06	0.30	0.60	0.208	0.957	0.242
WFDC12	CEU	6	2756	2.12	1.53	1.11	0.158	1.21	0.07	0.78	0.084	0.562	0.106
PI3	CEU	21	4546	6.54	5.33	0.80	0.228	1.04	0.07	-1.30	0.156	0.364	0.129
SEMG1	CEU	8	3964	2.09	2.02	0.10	0.344	0.69	0.12	-0.57	0.255	0.893	0.215
SEMG2	CEU	6	3658	1.71	1.52	0.35	0.339	1.21	0.06	-2.18	0.066	0.194	0.232
SLPI	CEU	10	5366	1.86	2.56	-0.88	0.227	-1.82	0.34	-2.60	0.040	0.679	0.505
WFDC2	CEU	6	3978	0.80	1.53	-1.38	0.088	-2.66	0.06	-1.68	0.102	0.969	0.757
SPINT3	CEU	4	3722	1.16	1.04	0.30	0.364	-1.08	0.27	-0.24	0.357	0.696	0.184
WFDC6	CEU	3	2803	0.61	0.80	-0.59	0.320	-1.48	0.13	0.49	0.285	0.596	0.552
EPPIN	CEU	3	3431	0.81	0.72	0.28	0.287	-0.32	0.42	0.71	0.094	0.872	0.267
WFDC8	CEU	27	7169	10.70	6.89	2.02	0.041	0.01	0.46	-0.25	0.406	0.143	0.009
WFDC9/10A	CEU	24	6855	7.93	6.10	1.07	0.186	0.87	0.12	-0.51	0.313	0.548	0.094
WFDC11	CEU	14	5032	2.93	3.56	-0.60	0.268	0.27	0.35	-1.70	0.112	0.136	0.479
WFDC10B/13	CEU	13	7356	3.70	3.30	0.40	0.244	-0.27	0.38	-0.13	0.391	0.848	0.130
SPINT4	CEU	9	3521	3.69	2.32	1.88	0.058	0.74	0.17	-1.06	0.206	0.109	0.061
WFDC3	CEU	31	7562	11.50	7.95	1.63	0.071	1.09	0.04	0.30	0.360	0.197	0.035
WFDC5	YRI	11	3787	2.64	2.47	0.21	0.116	0.29	0.17	0.77	0.136	0.762	0.140
WFDC12	YRI	14	2756	2.79	3.13	-0.33	0.351	-0.39	0.49	0.08	0.271	0.914	0.301
PI3	YRI	34	4546	4.70	7.60	-1.29	0.118	0.66	0.10	-1.71	0.016	0.409	0.833
SEMG1	YRI	20	3964	2.70	4.48	-1.27	0.242	0.21	0.18	0.12	0.296	0.655	0.760
SEMG2	YRI	18	3658	2.82	4.02	-0.95	0.375	0.84	0.06	-1.68	0.021	0.557	0.483
SLPI	YRI	29	5366	6.20	6.50	-0.15	0.182	0.17	0.19	-0.31	0.143	0.617	0.181
WFDC2	YRI	10	3978	2.49	2.23	0.34	0.125	0.78	0.11	0.10	0.277	0.945	0.085
SPINT3	YRI	11	3722	1.33	2.46	-1.34	0.136	-1.43	0.82	-1.60	0.022	0.234	0.809
WFDC6	YRI	12	2803	2.04	2.71	-0.75	0.449	0.94	0.05	0.13	0.302	0.541	0.352
EPPIN	YRI	17	3431	3.01	3.81	-0.67	0.350	0.79	0.07	0.33	0.479	0.075	0.242
WFDC8	YRI	38	7169	10.50	8.56	0.78	0.072	0.36	0.20	-0.88	0.067	0.300	0.061
WFDC9/10A	YRI	33	6855	8.85	7.40	0.66	0.090	0.87	0.06	0.35	0.412	0.861	0.054
WFDC11	YRI	18	5032	4.60	4.03	0.45	0.134	0.44	0.18	0.49	0.456	0.388	0.117

WFDC10B/13	YRI	37	7356	7.08	8.34	-0.52	0.271	0.77	0.04	0.43	0.489	0.230	0.150
SPINT4	YRI	18	3521	2.86	4.17	-1.03	0.172	-0.76	0.34	-0.62	0.093	0.084	0.706
WFDC3	YRI	55	7562	14.10	12.30	0.50	0.088	0.72	0.05	0.52	0.445	0.100	0.070

[a] S = number of segregating sites.

[b] Watterson's estimator of Θ ($4N_e\mu$) (Watterson 1975) per base pair (x 10^{-4}).

[c] Nucleotide diversity per base pair (x 10^{-4})

[d] Tajima's *D* statistic (Tajima 1989).

[e] Fu & Li *D** statistic (Fu and Li 1993)

[f] Corrected Fay and Wu *H* test (Fay and Wu 2002 and Zeng et al. 2006)

[g] P-value of HKA test (Hudson, Kreitman, Aguadé 1987)

[h] P-value of MWUhigh test (Andrés et al., 2009)

Supplementary Table S4: Summary statistics for WFDC locus genes in the 1000 genomes dataset, performed using SLIDER.

Gene	Population	Length	[a]S	[b]π (10^{-4})	[c]Θ_w	[d]Tajima's *D*	[e]Fu & Li *D**
WFDC5	CHB + JPT	3788	17	1.3421263	0.000691	-1.1935	-0.8464
WFDC12	CHB + JPT	2762	14	1.5648111	0.0007804	-0.6457	0.1111
PI3	CHB + JPT	4558	28	1.3402799	0.0009459	-1.8297	-1.6737
SEMG1	CHB + JPT	3972	25	0.8050052	0.0009691	-2.0666	-3.9752
SEMG2	CHB + JPT	3664	14	0.6341285	0.0005883	-1.6628	-1.9543
SLPI	CHB + JPT	5247	28	2.974262	0.0008216	-0.8233	0.6021
WFDC2	CHB + JPT	3990	15	3.4546218	0.0005788	1.185	-0.4528
SPINT3	CHB + JPT	3730	25	3.8286212	0.001032	-0.014	-0.0685
WFDC6	CHB + JPT	2811	13	1.9446779	0.0007121	-0.0659	-1.4313
EPPIN	CHB + JPT	3233	15	1.9801155	0.0007144	-0.3409	-0.4528
WFDC8	CHB + JPT	7189	44	4.3929822	0.0009424	-0.9829	0.6087
WFDC9/10A	CHB + JPT	6870	33	8.8730859	0.0007396	2.0234	0.8591
WFDC11	CHB + JPT	5042	22	4.5206572	0.0006718	0.8573	-1.9024
WFDC10B/12	CHB + JPT	7374	26	4.0535615	0.0005429	0.033	-2.8532
SPINT4	CHB + JPT	3530	33	10.84003	0.0014394	3.0729	0.449
WFDC3	CHB + JPT	7582	44	10.708893	0.0008935	1.6237	-1.4283
WFDC5	CEU	3788	14	2.8524398	0.0006446	0.4291	0.2845
WFDC12	CEU	2762	12	3.7390851	0.0007578	1.9447	1.4424
PI3	CEU	4558	27	4.8184897	0.0010332	0.0657	-0.5158
SEMG1	CEU	3972	13	1.6558871	0.0005709	-0.6781	-2.4283
SEMG2	CEU	3664	16	1.7599658	0.0007617	-0.9664	-2.9829
SLPI	CEU	5247	27	2.6329971	0.0008975	-1.2486	-0.1076
WFDC2	CEU	3990	13	1.2455184	0.0005683	-1.1332	-0.4697
SPINT3	CEU	3730	16	4.0205297	0.0007482	1.1531	-1.2598
WFDC6	CEU	2811	6	1.0904993	0.0003723	0.087	-1.9902
EPPIN	CEU	3233	6	1.0656938	0.0003237	0.038	0.0537
WFDC8	CEU	7189	43	12.826326	0.0010433	2.1116	1.0398

WFDC9/10A	CEU	6870	40	11.834331	0.0010155	2.057	1.8619
WFDC11	CEU	5042	18	4.6422702	0.0006227	1.2782	0.6173
WFDC10B/12	CEU	3530	35	10.076526	0.0017293	1.8975	-0.293
SPINT4	CEU	7374	24	4.6894216	0.0005677	0.3354	0.0929
WFDC3	CEU	7582	41	9.3882599	0.0009432	0.9263	-0.8532
WFDC5	YRI	3788	19	3.2246865	0.0008731	-0.0672	-0.3477
WFDC12	YRI	2762	20	3.3670713	0.0012605	-0.0891	1.7482
PI3	YRI	4558	32	3.9328703	0.0012221	-0.8482	1.3107
SEMG1	YRI	3972	26	2.2172896	0.0011394	-1.436	-0.1882
SEMG2	YRI	3664	20	2.3281056	0.0009502	-0.8986	0.2503
SLPI	YRI	5247	36	7.6492865	0.0011943	0.6447	1.7751
WFDC2	YRI	3990	21	3.6158577	0.0009162	-0.0296	0.8109
SPINT3	YRI	3730	32	5.2212913	0.0014934	-0.1807	0.5829
WFDC6	YRI	2811	18	2.797198	0.0011146	-0.286	-0.4535
EPPIN	YRI	3233	25	3.3676545	0.001346	-0.6334	0.1623
WFDC8	YRI	7189	62	13.967005	0.0015012	0.8978	-0.0326
WFDC9/10A	YRI	6870	47	11.791325	0.0011909	1.3204	1.4403
WFDC11	YRI	5042	21	5.4833263	0.000725	1.3665	0.8109
WFDC10B/13	YRI	7374	53	7.1873641	0.0012511	-0.6672	0.3216
SPINT4	YRI	3530	40	7.7484559	0.0019725	0.333	1.242
WFDC3	YRI	7582	52	11.871137	0.0011938	0.9394	1.3027

[a] S = number of segregating sites.

[b] Watterson's estimator of Θ ($4N_e\mu$) (Watterson 1975) per base pair (x 10^{-4}).

[c] Nucleotide diversity per base pair (x 10^{-4})

[d] Tajima's *D* statistic (Tajima 1989).

[e] Fu & Li *D** statistic (Fu and Li 1993)

Supplementary Table S5: Independence (χ^2) and correlation (Kendall's τ and Spearman's ρ) tests between the equivalent summary statistics distributions of the 1000 genomes and the sequenced data.

	π (10^{-4})	Θ_w	Tajima's *D*	Fu & Li *D**
χ^2	0.025	0.902	0.999	0.159
Kendall's τ	0.674	0.654	0.619	0.150
Spearman's ρ	0.850	0.829	0.808	0.216

Supplementary Table S6: Command lines used in the ms program to test Gutenkunst model.

ms command line without outgroup:
ms 142 REPETITIONS -s SEGSITES RECOMBINATION_RATE -I 3 50 42 50 -n 1 1.682020 -n 2 3.736830 -n 3 7.292050 -eg 0 2 116.010723 -eg 0 3 160.246047 -ma x 0.881098 0.561966 0.881098 x 2.797460 0.561966 2.797460 x -ej 0.028985 3 2 -en

0.028985 2 0.287184 -ema 0.028985 3 x 7.293140 x 7.293140 x x x x -ej 0.197963 2 1 -en 0.303501 1 1

ms command line with outgroup:

ms 143 REPETITIONS -s SEGSITES RECOMBINATION_RATE -I 4 50 42 50 1 -n 1 1.682020 -n 2 3.736830 -n 3 7.292050 -n 4 1.3699 -eg 0 2 116.010723 -eg 0 3 160.246047 -ma x 0.881098 0.561966 0 0.881098 x 2.797460 0 0.561966 2.797460 x 0 0 0 0 x -ej 0.028985 3 2 -en 0.028985 2 0.287184 -ema 0.028985 4 x 7.293140 x 0 7.293140 x x 0 x x x x 0 0 x x -ej 0.197963 2 1 -en 0.303501 1 1 -em 16.4300 1 4 0 -em 16.4300 4 1 0 -ej 3.69492340746 4 1

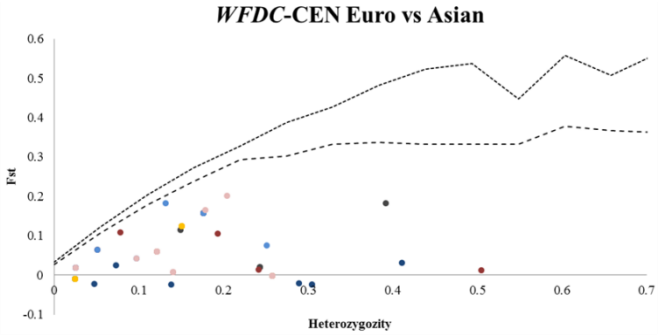
REPETITIONS = 10.000

SEGSITES = Segregating sites calculated for each population for each gene independently.

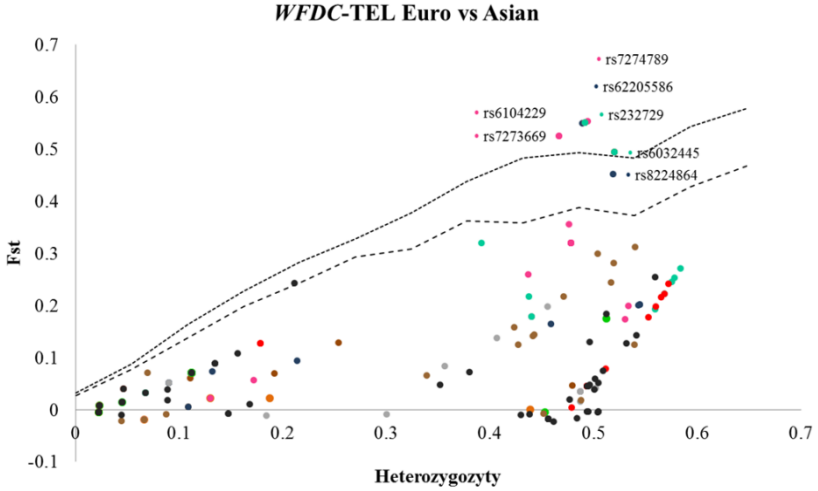
RECOMBINATION RATE = Calculated using the UCSC browser, 0.20 cM/Mb, in the *WFDC* region.

Supplementary Figures

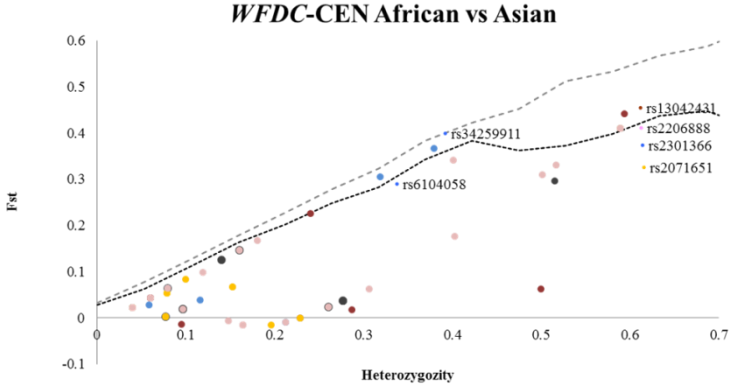
A)



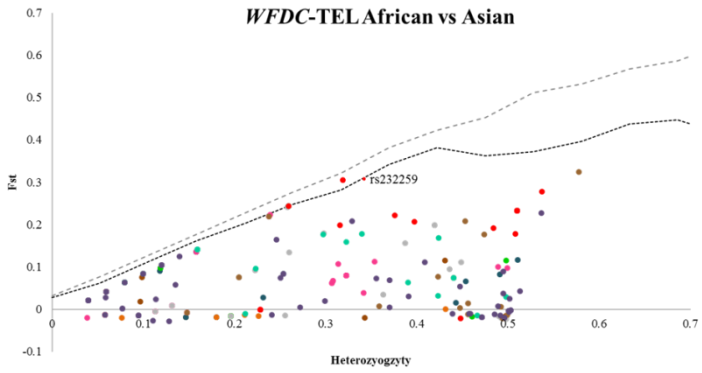
B)

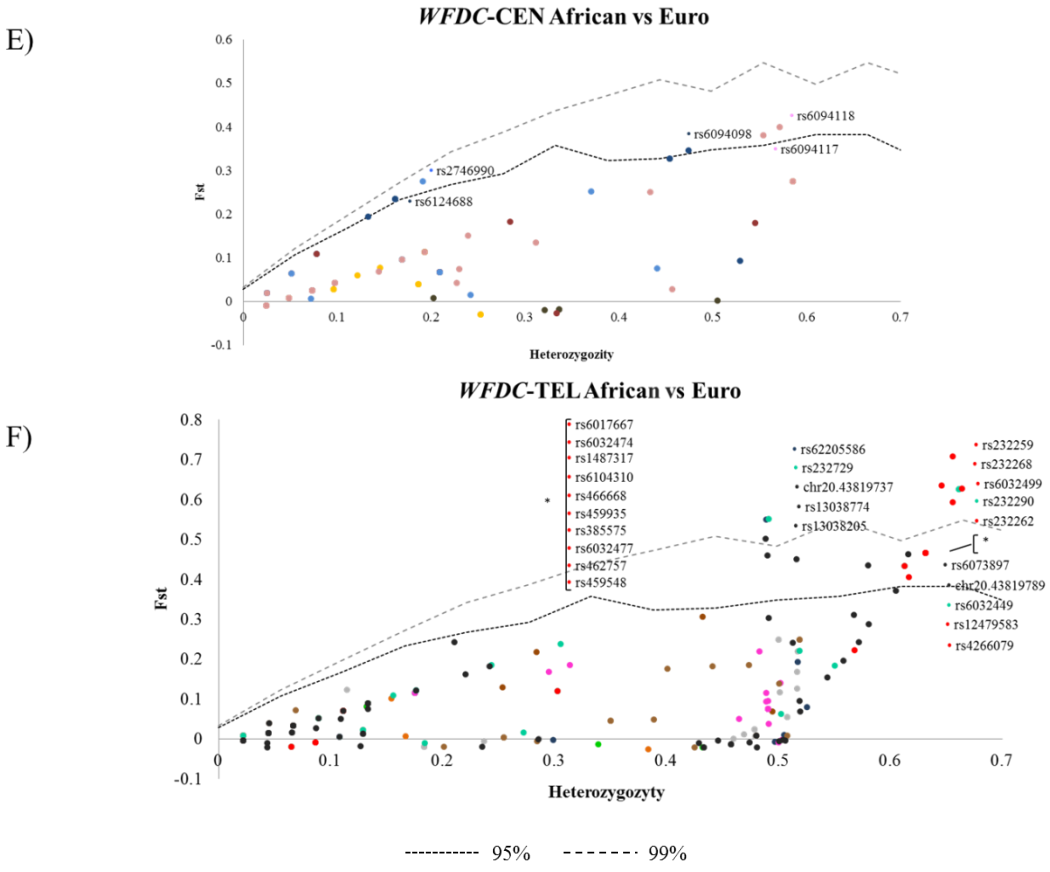


C)

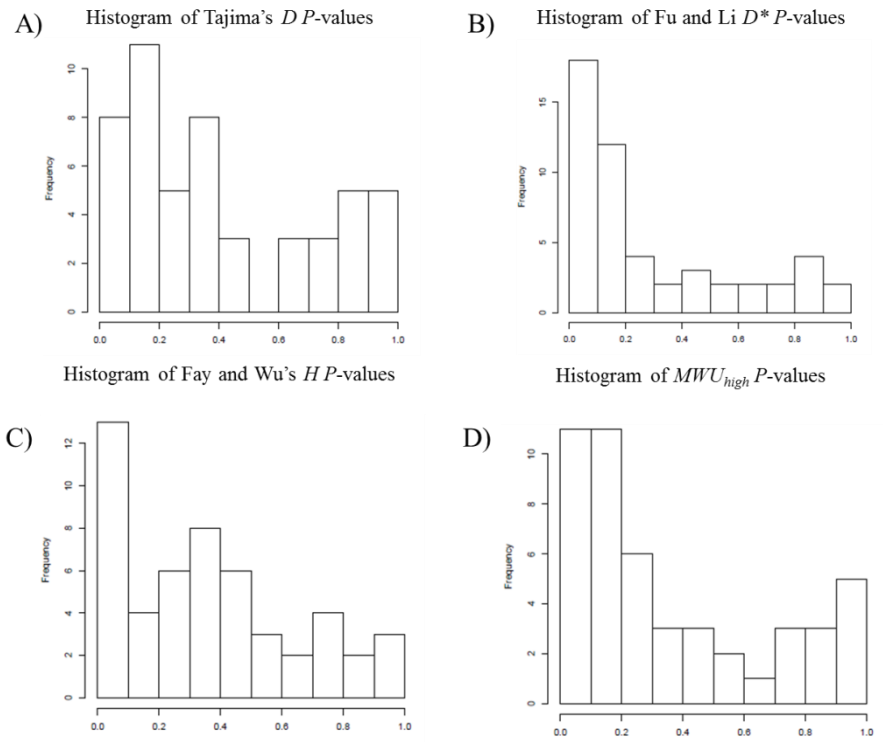


D)

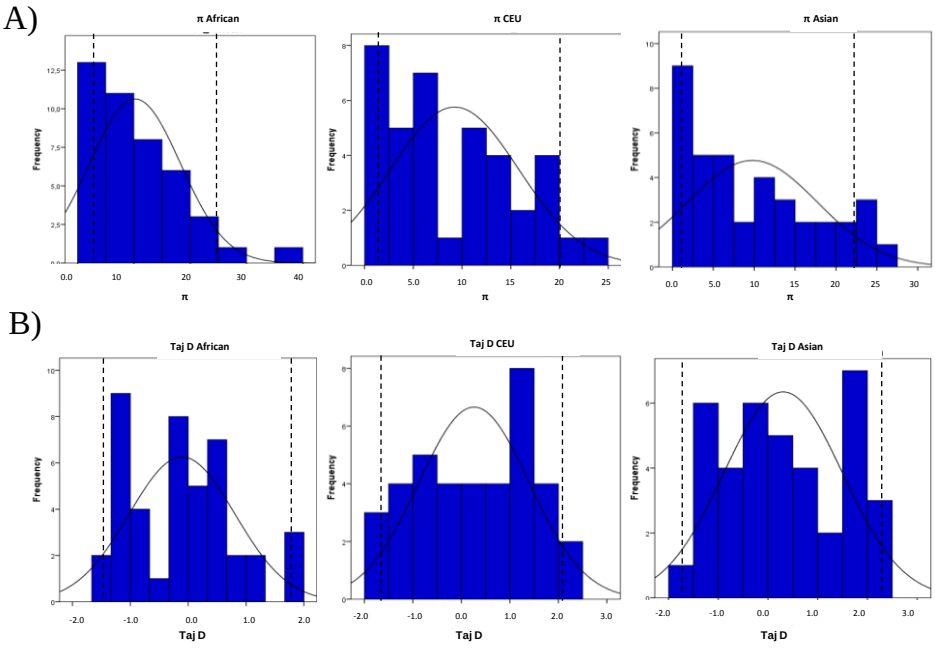




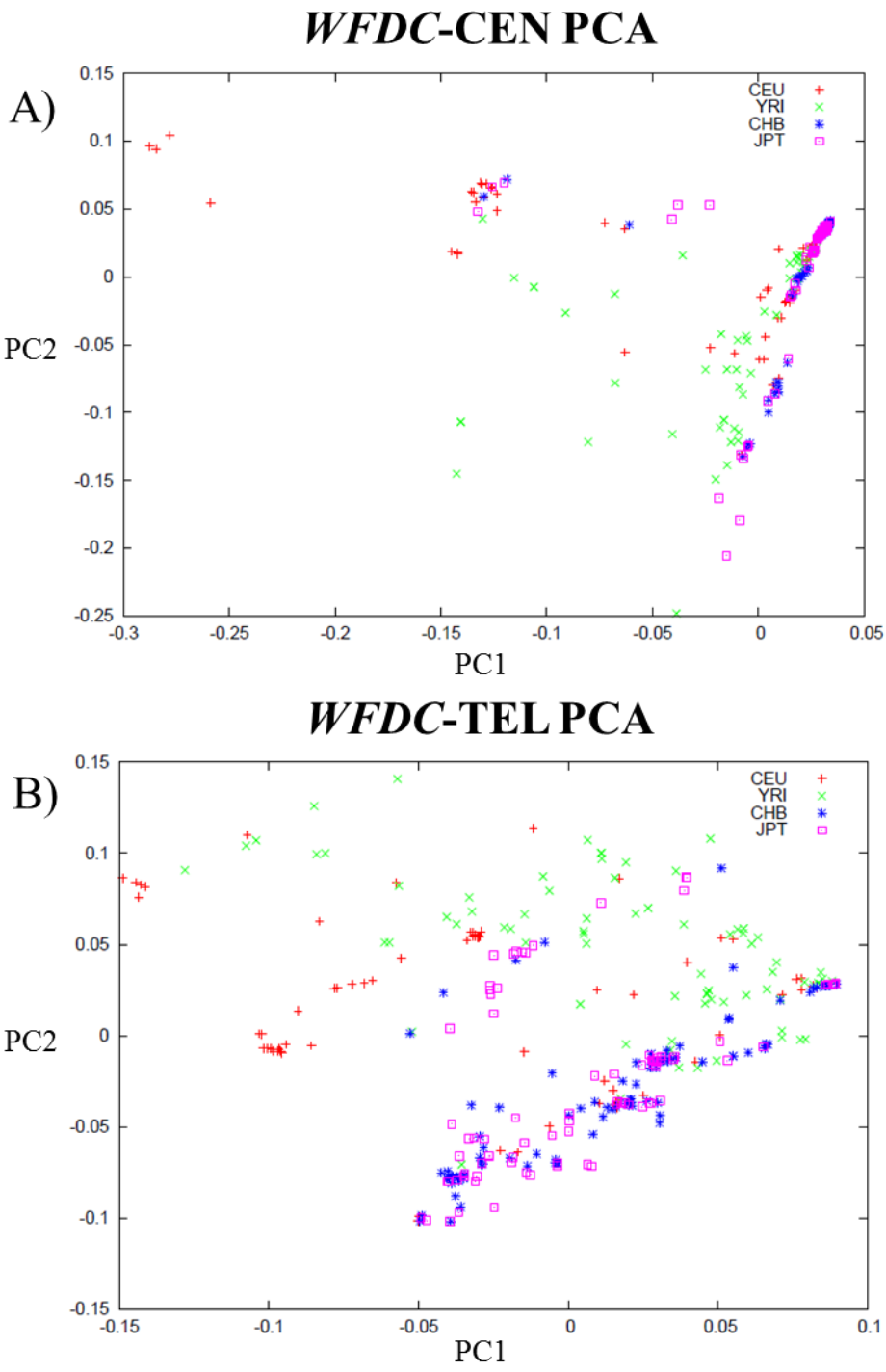
Supplementary Figure S1: F_{ST} statistics (Excoffier 2002) used to describe the proportion of genetic variance attributable to between-population (European *vs*. Asian, African vs European, and African vs Asian) effects in the *WFDC*-CEN (A), C) and E)) and *WFDC*-TEL (B), D), and F)). To identify SNPs presenting extreme levels of F_{ST}, we compared the observed F_{ST} at each SNP within the *WFDC* region with a locus-by-locus AMOVA of the control regions (10,000 permutations) using 20,000 simulations, performed by Arlequin (Excoffier, Laval, Schneider 2005).



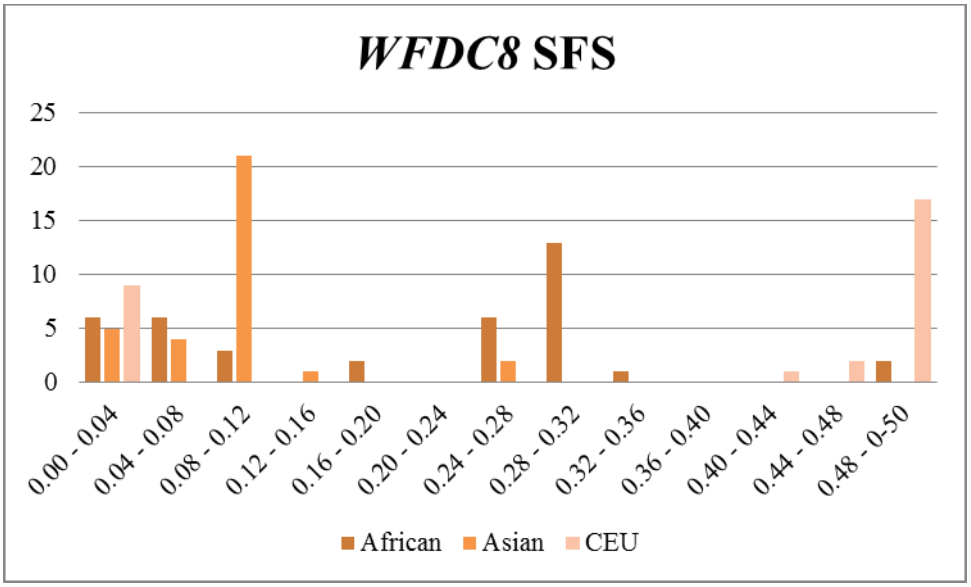
Supplementary Figure S2: Histogram representation of A) Tajima's *D*, B) Fu and Li's *D*, C) Fay and Wu's *H* and D) MWU_{high} *P*-values, showing a clear augmentation of low *P*-values.



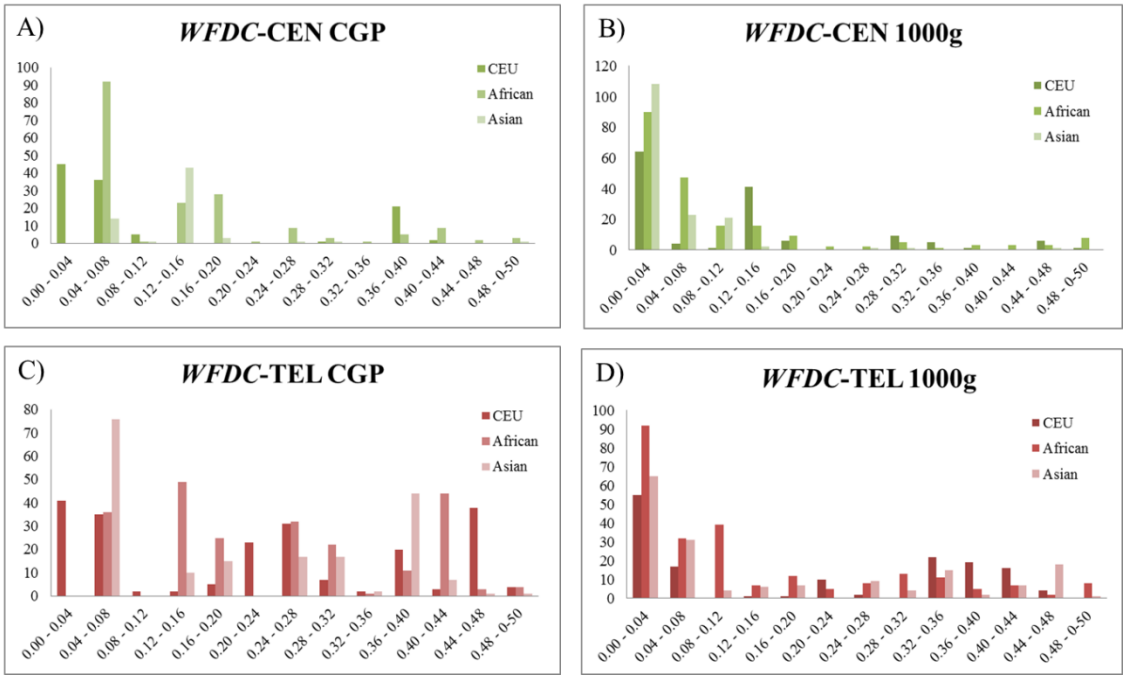
Supplementary Figure S3: Empirical distributions of A) π and B) Tajima's *D*, determined from the neutral control genomic regions in each population. Dashed lines represent upper and lower 2.5 percentiles, calculated using the SPSS statistics software package, version 20 (IBM Corporation: Chicago, IL, USA, 2010).



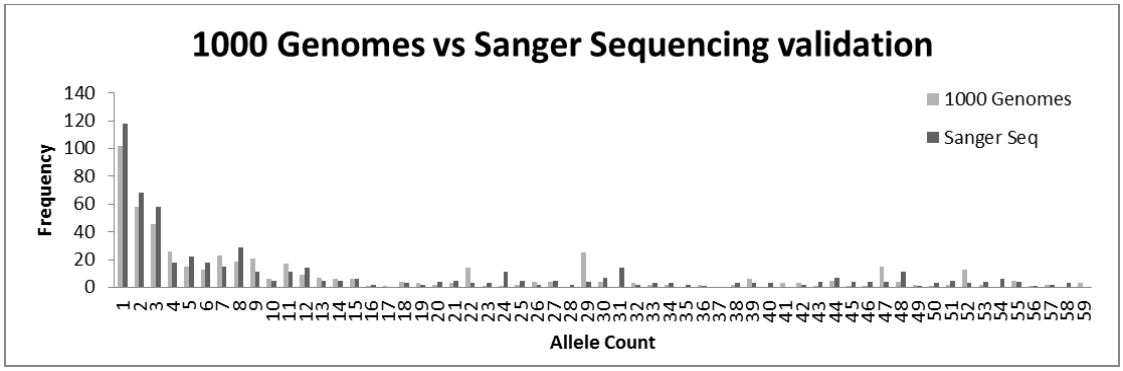
Supplementary Figure S4: The corresponding principal components plots for the sequences of the centromeric A) and telomeric B) sub-loci downloaded from the 1000 Genomes Project. A Principal Component Analysis (PCA) was performed using EigenSoft (Patterson 2006). 80% of the variation is explained by PC1.



Supplementary Figure S5: *WFDC8* site frequency spectrum in the three different sequenced populations (Africa, Asian and European).



Supplementary Figure S6: Site Frequency Spectrum (SFS) of the corresponding centromeric A) and C) and telomeric B) and D) sub-loci data downloaded from 1000 genomes and from the Complete Genomics Diversity Panel.



Supplementary Figure S7: Allele Frequency Spectrum of the same regions and individuals sequenced by Sanger sequencing and in the 1000 genomes project dataset.

Appendix III

Supplementary Material Article 3

Sequence diversity of *Pan troglodytes* subspecies and the impact of *WFDC6* selective constraints in reproductive immunity.

(submitted)

Supplementary Tables

Supplementary Table S1: Regions of the genome sequenced.

WFDC locus region

Chr 20 coordinates*	Gene
	WFDC-CEN
43,171,507-43,177,217	*WFDC5* (WAP four-disulfide core domain 5)
43,185,481-43,186,520	*WFDC12* (WAP four-disulfide core domain 12)
43,236,912-43,238,599	*PI3* (elafin)
43,269,088-43,271,823	*SEMG1* (semenogelin 1)
43,269,088-43,286,513	*SEMG2* (semenogelin 2)
43,314,293-43,316,620	*SLPI* (secretory leukocyte peptidase inhibitor)
	WFDC-TEL
43,541,899-43,543,586	*WFDC2* (WAP four-disulfide core domain 2)
43,574,515-43,577,678	*SPINT3*
43,596,250-43,601,548	*WFDC6* (WAP four-disulfide core domain 6)
43,602,679-43,609,442	*SPINLW1* (WAP four-disulfide core domain 7)
43,613,815-43,641,379	*WFDC8* (WAP four-disulfide core domain 8)
43,669,992-43,693,321	*WFDC9* (WAP four-disulfide core domain 9)
43,691,799-43,693,245	*WFDC10a* (WAP four-disulfide core domain 10a)
43,710,616-43,732,292	*WFDC11* (WAP four-disulfide core domain 11)
43,746,704-43,767,072	*WFDC10b* (WAP four-disulfide core domain 10b)
43,764,069-43,770,870	*WFDC13* (WAP four-disulfide core domain 13)
43,784,402-43,787,749	*SPINT4*
43,836,254-43,853,954	*WFDC3* (WAP four-disulfide core domain 3)

#Control Regions (from Andres, 2010)

Ψ ID	Chrromosome coordinate	Gene and chr of origin			Processed Ψs per genome			
					Human	Chimp	Oran	Rhesus
ENCODE Ψ:79794	chr11:5505093-5505962	Unknown			1	1	1	1
NON-ENCODE Ψ:89	chr1:38020719-38021144	C14ORF138	@	Chr14	1	1	1	0
Ψ:127271	chr1:224666087-224666538	Unknown			1	3	2	4
Ψ:511	chr1:181191136-181191557	NCR1	@	Chr19	1	1	1	1
Ψ:716	chr1:245414326-245414818	PEX5	@	Chr12	1	1	1	1
Ψ:189	chr1:61892410-61893166	SPR	@	Chr2	1	1	1	1
Ψ:414	chr1:156412265-156414162	ELL2	@	Chr5	3	3	3	3
Ψ:725	chr2:7794021-7794677	PSMB1	@	Chr6	1	1	ORF	2?
Ψ:881	chr2:76338812-76339387	NP	@	Chr14	1	1	1	1 + ORF
Ψ:919	chr2:101503725-101504888	PRCP	@	Chr11	1	1	1	1
Ψ:10377	chr2:146951393-146952148	Unknown			1	1	1	1
Ψ:10800	chr2:158526147-158527646	MTA3	@	Chr2	1	1	1	ND

Ψ:1317	chr3:34889168-34890350	FECH	@	Chr18	1	1	1	1
Ψ:1390	chr3:59496630-59497181	Unknown			2	2	2	2
Ψ:131502	chr3:81498830-81499811	Unknown			1	1	1	ND
Ψ:1588	chr3:146885050-146885571	GM2A	@	Chr5	2	2	2	2
Ψ:75498	chr4:36184061-36186883	FBXO38	@	Chr5	1	1	1	1
Ψ:1886	chr4:89667737-89668341	CD53	@	Chr1	1	1	1	1
Ψ:18262	chr4:176732748-176734338	ADAM29	@	Chr4	2	2	2	2
Ψ:1835	chr4:68730627-68731081	Unknown			1	1	1	1
Ψ:2117	chr5:29141432-29142101	C14ORF45	@	Chr14	1	1	1	1
Ψ:19453	chr5:133084092-133084524	DPH4	@	Chr11	1	1	1	1
Ψ:2934	chr6:139701013-139702410	DNAJC7	@	Chr17	2	2	2	3
Ψ:3153	chr7:55279435-55280868	SLC19A3	@	Chr2	2	2	2	2
Ψ:20740	chr7:128120741-128121217	IMP3	@	Chr15	2	2	2	2
Ψ:3607	chr8:60632574-60632998	NUDT15	@	Chr13	2	2 + 1 ORF	2	2
Ψ:3693	chr8:97208019-97209206	Unknown			1	1	1	1
Ψ:69522	chr9:6185935-6186853	GTF3A	@	Chr13	5	5	5	5
Ψ:21854	chr9:21685170-21686842	KHSRP	@	Chr19	1	1	1	1
Ψ:4067	chr9:106130426-106131316	WDR45L	@	Chr17	3	2 + 1 ORF	2 + 1 ORF	2
Ψ:4320	chr10:79498396-79499304	GNAI2	@	Chr3	3	3	2	1
Ψ:4547	chr11:40061840-40062643	ZCCHC9	@	Chr5	3	2	3	3
Ψ:4837	chr11:113829007-113830057	CCRN4L	@	Chr4	1	1	1	2
Ψ:4765	chr11:91708853-91709310	NDUFB11	@	ChrX	1	1	1	1 + 1 ORF
Ψ:4913	chr12:7650698-7651196	CLTA	@	Chr9	2	2	2	2
Ψ:5244	chr12:107199014-107200142	APOBEC3F	@	Chr22	1	1	1	1
Ψ:5055	chr12:44124847-44125485	MESDC2	@	Chr15	1	1	1	1
Ψ:73106	chr13:24217272-24218374	IRX1	@	Chr5	1	1	1	1
Ψ:5504	chr13:89442087-89443147	PEX12	@	Chr17	1	1	1	0
Ψ:5477	chr13:74300713-74302220	RIOK3	@	Chr18	1	1	1	1
Ψ:5700	chr14:61368336-61368840	COX4I1	@	Chr16	2	1	2	2 + 1 ORF
Ψ:62081	chr14:88646206-88647320	MPPE1	@	Chr18	2	1	1	1
Ψ:6310	chr16:82710324-82711542	PLK-1	@	Chr16	2	2	2	2
Ψ:6549	chr17:66136577-66137001	Unknown			1	1	1	1
Ψ:26649	chr19:18012148-18012866	APOA1BP	@	Chr1	2	2	2	2
Ψ:6997	chr20:21429964-21430620	GSTM3	@	Chr1	1	1	1	2
Ψ:7045	chr20:37391075-37391973	ATG3	@	Chr3	1	2	1	1
Ψ:7112	chr21:15051781-15052541	POLR2C	@	Chr16	1	1	1	1
Ψ:7176	chr21:45316169-45316649	Unknown			1	1	1	1

* Chromosome coordinates are based in the human march 2006 assembly (hg18; genome.ucsc.edu)

[#] These loci were selected using the following filters: Processed ψs (according to the annotation in ψ.org); Minimum length of 400 bp; Unlinked to each other; not ribosomal RNA (according to the annotation in ψ.org); With human, orangutan and rhesus orthologs; No overlap with UCSC genes; No overlap with highly conserved regions (most conserved, UCSC); Average genome recombination rate; Processed ψs in single copy in human and chimpanzee genomes were preferred. (When this was not possible, one ψ per family was accepted provided that the members of the family had 90% identity or less among them.)

Supplementary Table S2: **Chimpanzee samples sequenced.**

Population	**Sample numbers**						
Pan troglodytes troglodytes (15)	Etrange	Ewake		Kita	Mac	Maya	Mokolo	
		Nanga-Eboko		Nikita	Papaie	Pauldina	Pecos	Silva
		Suzanne		Lada	Jules			
Pan troglodytes ellioti (31)	Achidi		Akwaya-Jean		Alex	Bally	Bankim	Basho
			Bernadette		Carlos	Eve	Gah	
		George	Jack		Jacob	Julie	Kopongo	
		Louisa	Margaret		Mesange		Nemo	
		Nicolene	Papa	Paquita		TKC	Taweh	Tinto-
	Mbu	Tobi	Bana	Bergkamp		Koto	Damian	
		Banyo						
Pan troglodytes verus (22)		Pt100	Pt101	Pt102	Pt103	Pt104	Pt105	Pt106
			Pt107	Pt112	Pt115	Pt117	Pt120	Pt121
		Pt122	Pt124	Pt125	Pt126	Pt81	Pt87	Pt88
		Pt97	Pt98					

Supplementary Table S3: Summary Statistics for all the *WFDC* genes.

Gene	Sub-species	S	π (10^{-4})	Length	Θ_w	Tajima's *D*	Fu & Li *D**	Fay and Wu's *H*	P(HKA)
WFDC5	*P. t. troglodytes*	32	11.87	5536	8.077	-0.4202	-0.3432	-1.6644	0.3858
WFDC12	*P. t. troglodytes*	16	3.026	1323	4.039	-0.6318	0.4407	-0.6437	0.1866
PI3	*P. t. troglodytes*	25	7.748	3377	6.310	-0.3775	0.1365	1.5448	0.9625
SEMG1	*P. t. troglodytes*	23	4.015	3305	5.806	-1.172	-0.8307	-1.2782	0.7463
SEMG2	*P. t. troglodytes*	31	8.471	4324	7.825	-0.8778	0.2350	-2.4092	0.9241
SLPI	*P. t. troglodytes*	30	16.84	4709	7.573	0.5078	0.6260	2.5195	0.9095
WFDC2	*P. t. troglodytes*	24	5.850	3984	6.058	-0.7272	0.3391	-3.6322	0.4648
SPINT3	*P. t. troglodytes*	35	10.72	3727	8.835	-0.8613	-0.9377	14.278	0.0604
WFDC6	*P. t. troglodytes*	19	1.269	2807	4.796	-2.073	-2.290	-1.7839	0.1907
WFDC7	*P. t. troglodytes*	23	2.356	3233	5.806	-1.811	-2.218	5.2828	0.1362
WFDC8	*P. t. troglodytes*	36	9.139	7179	9.087	-1.169	-1.054	0.5977	0.5583
WFDC9/10A	*P. t. troglodytes*	44	16.64	6863	11.11	-0.8419	-0.8662	2.0736	0.2643
WFDC11	*P. t. troglodytes*	58	37.48	5037	14.64	-0.3621	-0.1044	9.0184	0.4493
WFDC10B/13	*P. t. troglodytes*	41	13.99	7365	10.35	-0.9045	-0.8869	-5.0023	0.882
SPINT4	*P. t. troglodytes*	14	2.296	3527	3.534	-0.7060	-0.1432	-1.9862	0.9805
WFDC3	*P. t. troglodytes*	51	20.48	7572	12.87	-0.9501	-0.6436	-0.5163	0.4458
WFDC5	*P. t.ellioti*	24	9.751	5536	5.110	0.8670	0.4099	-1.9799	0.1653
WFDC12	*P. t.ellioti*	7	1.277	1323	1.491	0.8178	0.4173	-0.0063	0.9754
PI3	*P. t.ellioti*	14	3.947	3377	2.981	0.9111	-0.4956	1.1021	0.3997
SEMG1	*P. t.ellioti*	28	3.491	3305	5.962	-1.250	-0.5904	-0.9815	0.321
SEMG2	*P. t.ellioti*	22	2.378	4324	4.685	-1.193	-0.4544	-0.1861	0.2107
SLPI	*P. t.ellioti*	30	5.464	4709	6.388	-0.8478	-1.875	4.6113	0.9302
SPINT3	*P. t.ellioti*	14	3.866	3727	2.981	0.8654	0.5262	0.7953	0.2557
WFDC2	*P. t.ellioti*	25	7.905	3984	5.323	0.2832	-0.1960	5.2226	0.5012
WFDC6	*P. t.ellioti*	13	1.364	2807	2.768	-0.7455	-1.177	-1.6838	0.3477
EPPIN	*P. t.ellioti*	23	3.936	3233	4.898	-0.6386	-0.0079	12.0571	0.054
WFDC8	*P. t.ellioti*	29	5.759	7179	6.175	-0.6888	-1.990	4.7842	0.3548
WFDC9 10A	*P. t.ellioti*	28	10.18	6863	5.962	0.3822	0.9311	3.1793	0.0392
WFDC11	*P. t.ellioti*	42	15.62	5037	8.943	-0.1915	-0.1990	4.7848	0.2656
WFDC10B	*P. t.ellioti*	29	6.956	7365	6.175	-0.4046	-0.2142	0.055	0.2968
SPINT4	*P. t.ellioti*	14	0.8043	3527	2.981	-1.498	-1.007	4.3977	0.8262
WFDC3	*P. t.ellioti*	59	14.79	7572	12.56	-1.182	-0.8554	15.3739	0.7554
WFDC5	*P. t. verus*	14	1.743	5536	3.218	-0.8086	0.6041	9.148	0.4733
WFDC12	*P. t. verus*	7	1.396	1323	1.609	0.7787	-1.027	-0.8393	0.1127
PI3	*P. t. verus*	7	1.979	3377	1.609	1.6256	0.4908	0.8076	0.6147
SEMG1	*P. t. verus*	20	2.323	3305	4.598	-1.2464	-0.8163	-1.4884	0.0296
SEMG2	*P. t. verus*	9	0.883	4324	2.069	-0.7344	-1.207	-0.5666	0.2473

SLPI	*P. t. verus*	14	2.797	4709	3.218	-0.0454	1.071	0.8203	0.8173
WFDC2	*P. t. verus*	3	0.212	3984	0.690	-0.4387	-0.3775	0.7653	0.3611
SPINT3	*P. t. verus*	9	1.865	3727	2.069	0.5711	0.0750	3.3425	0.6688
WFDC6	*P. t. verus*	9	0.629	2807	2.069	-1.1725	-0.5657	-1.0973	0.0616
WFDC7	*P. t. verus*	7	1.523	3233	1.609	0.9748	-0.2682	-0.4207	0.2532
WFDC8	*P. t. verus*	7	0.815	7179	1.609	-0.2537	0.4908	-1.1501	0.0954
WFDC9/10A	*P. t. verus*	22	3.333	6863	5.057	-1.002	-0.2720	3.2896	0.4979
WFDC11	*P. t. verus*	25	7.986	5037	5.747	0.0283	-0.3329	2.408	0.0605
WFDC10B/13	*P. t. verus*	14	3.356	7365	3.218	0.3020	1.071	0.4249	0.4759
SPINT4	*P. t. verus*	6	1.062	3527	1.379	0.6795	-0.4940	-0.4989	0.856
WFDC3	*P. t. verus*	32	17.27	7572	7.356	0.6888	0.3704	12.222	0.8309
WFDC5	*Pan troglodytes*	45	7.7796	5536	8.25	-0.1739	-0.1817	1.4109	0.6746
WFDC12	*Pan troglodytes*	24	2.4808	1323	4.39	-1.2459	-1.0808	0.7227	0.6637
PI3	*Pan troglodytes*	32	4.3176	3377	5.85	-0.7764	-0.3782	6.5226	0.4454
SEMG1	*Pan troglodytes*	45	3.9938	3305	8.2	-1.5677	0.3357	-1.0394	0.1462
SEMG2	*Pan troglodytes*	47	3.5766	4324	8.57	-1.7859	-0.352	5.0682	0.8163
SLPI	*Pan troglodytes*	46	6.6832	4709	8.38	-0.6211	0.6394	7.6619	0.247
WFDC2	*Pan troglodytes*	34	5.3535	3984	6.2	-0.4058	-0.2482	0.656	0.1038
SPINT3	*Pan troglodytes*	44	5.2971	3727	8.04	-1.0425	-1.5881	36.879	0.0084
WFDC6	*Pan troglodytes*	31	1.6333	2807	5.67	-2.1039	-2.8978	-1.5222	0.0131
WFDC7	*Pan troglodytes*	38	3.165	3233	6.96	-1.648	-2.0926	16.8892	0.0175
WFDC8	*Pan troglodytes*	56	4.8159	7179	10.2	-1.6397	-2.1769	6.131	0.1153
WFDC9/10A	*Pan troglodytes*	67	7.2409	6863	12.2	-1.2844	-1.8648	19.3656	0.3861
WFDC11	*Pan troglodytes*	71	10.094	5037	13	-0.7082	-0.5289	25.782	0.0145
WFDC10B/13	*Pan troglodytes*	59	6.8495	7365	10.8	-1.1309	-1.54	-4.1285	0.1052
SPINT4	*Pan troglodytes*	30	2.2099	3527	5.5	-1.7644	-1.9467	16.0833	0.1734
WFDC3	*Pan troglodytes*	93	10.64	7572	17	-1.1874	-1.0211	17.8464	0.0719

Supplementary Table S4: Nonsynonymous substitutions *WFDC* genes.

a) Nonsynonymous chimpanzee variable positions

SNP location in chr20 (Pantro2)	Protein	Frequency	Expected Residue	Found Residue	Expected Codon	Found Codon	SIFT	PolyPhen
42453065	WFDC5	0.06	R	H	CGC	CAC	Tolerated	Benign
42453208	WFDC5	0.04	V	I	GTC	ATC	Tolerated	Possibly Damaging
42453271	WFDC5	0.02	V	M	GTG	ATG	Tolerated	Possibly Damaging
42466176	WFDC12	0.04	W	R	TGG	CGG	Damaging	Benign
42466200	WFDC12	0.04	S	P	TCA	CCA	Damaging	Benign
42466206	WFDC12	0.04	D	H	GAT	CAT	Damaging	Probably Damaging
42466400	WFDC12	0.01	K	N	AAG	AAT	Tolerated	Probably Damaging
42466463	WFDC12	0.02	D	E	GAC	GAA	Tolerated	Probably Damaging
42466492	WFDC12	0.02	D	N	GAT	AAT	Tolerated	Benign
42466510	WFDC12	0.04	V	I	GTA	ATA	Tolerated	Benign
42519166	PI3	0.02	T	M	ACG	ATG	Tolerated	Benign
42520083	PI3	0.12	V	I	GTT	ATT	Tolerated	Possibly Damaging
42568115	SEMG2	0.02	G	D	GGT	GAT	Affect Protein Funcion	Probably Damaging
42568172	SEMG2	0.05	G	V	GGC	GTC	Tolerated	Possibly Damaging
42568204	SEMG2	0.02	H	D	CAT	GAT	Tolerated	Benign
42568441	SEMG2	0.02	A	T	GCT	ACT	Tolerated	Possibly Damaging
42568613	SEMG2	0.01	S	N	AGC	AAC	Tolerated	Possibly Damaging
42568780	SEMG2	0.01	H	Y	CAT	TAT	Tolerated	Probably Damaging
42568831	SEMG2	0.02	K	E	AAG	GAG	Tolerated	Probably Damaging
42568883	SEMG2	0.02	K	M	AAG	ATG	Tolerated	Probably Damagin
42568888	SEMG2	0.02	H	Y	CAT	TAT	Tolerated	Probably Damaging
42569052	SEMG2	0.01	E	D	GAG	GAT	Tolerated	Probably Damaging
42569213	SEMG2	0.07	I	T	ATT	ACT	Tolerated	Benign
42569468	SEMG2	0.02	R	P	CGA	CCA	Tolerated	Probably Damaging
42569698	SEMG2	0.02	H	Y	CAT	TAT	Tolerated	Benign
42826868	WFDC2	0.02	S	L	TCG	TTG	Tolerated	Possibly Damaging
42826905	WFDC2	0.04	S	R	AGC	AGG	Tolerated	Benign
42897365	WFDC6	0.01	C	R	TGT	CGT	Tolerated	Probably Damaging
42897376	WFDC6	0.02	R	H	CGT	CAT	Tolerated	Benign
42897403	WFDC6	0.03	E	G	GAA	GGA	Tolerated	Benign
42897449	WFDC6	0.01	E	K	GAA	AAA	Tolerated	Benign
42897457	WFDC6	0.07	V	G	GTG	GGG	Tolerated	Probably Damaging
42898710	WFDC6	0.01	I	V	ATC	GTC	Tolerated	Benign
42972808	WFDC9	0.03	I	T	ATT	ACT	Torerated	Possibly Damaging

42992587	WFDC10A	0.09	T	P	ACT	CCT	Affect Protein Funcion	Benign
42992632	WFDC10A	0.01	Q	K	CAG	AAG	Torelated	Possibly Damaging
42993698	WFDC10A	0.01	C	R	TGT	CGT	Affect Protein Funcion	Probably Damaging
43044472	WFDC10B	0.01	C	*	TGT	TGA		
43044552	WFDC10B	0.02	L	V	CTA	GTA	Torelated	Benign
43044585	WFDC10B	0.01	I	V	ATC	GTC	Affect Protein Funcion	Benign
43045539	WFDC10B	0.07	R	C	CGT	TGT	Tolerated	Possibly Damaging
43045602	WFDC10B	0.08	P	T	CCC	ACC	Tolerated	Benign
43071581	WFDC10B	0.02	R	T	AGG	ACG	Affect Protein Function	Probably Damaging
43142946	WFDC3	0.01	E	K	GAA	AAA	Tolerated	Benign
43142988	WFDC3	0.02	K	E	AAA	GAA	Tolerated	Benign
43143039	WFDC3	0.02	C	S	TGT	AGT	Affect Protein Function	Probably Damaging
43155352	WFDC3	0.44	S	F	TCT	TTT	Tolerated	Benign
43155425	WFDC3	0.45	T	P	ACT	CCT	Tolerated	Benign
43155466	WFDC3	0.20	P	L	CCT	CTT	Tolerated	Benign
43155467	WFDC3	0.03	P	S	CCT	TCT	Affect Protein Function	Benign

b) Non-synonymous chimpanzee specific fixed differences

Protein	Position	Human	Chimp	Orang	Macaque	SIFT	PolyPhen
WFDC12	G27D	C	T	C	C	Tolerated	Benign
SEMG2	A93V	C	T	C	C	Tolerated	Benign
SEMG2	G101D	G	A	G	G	Tolerated	Benign
SEMG2	K120E	A	G	A	A	Tolerated	Benign
SEMG2	H136Y	C	T	C	C	Tolerated	Benign
SEMG2	S232G	A	G	A	A	Tolerated	Benign
SEMG2	H401R	A	G	A	A	Tolerated	Benign
SEMG2	H461R	A	G	A	A	Tolerated	Benign
SEMG2	T485S	A	T	A	A	Affect Protein Function	Benign
SEMG2	G504D	G	A	G	G	Tolerated	Benign
SLPI	L13F	G	A	G	G	Tolerated	Benign
WFDC6	P34S	G	A	G	G	Tolerated	Benign
EPPIN	K79E					Affect Protein Function	
WFDC8	L207P	A	G	?	A	Tolerated	Benign
WFDC8	D64E	G	T	G	A	Tolerated	Benign
WFDC8	S30C	G	C	G	G	Tolerated	Benign
WFDC9	F31L	A	G	A	T	Tolerated	Benign
WFDC10B/13	L16S	A	G	A	A	Tolerated	Benign

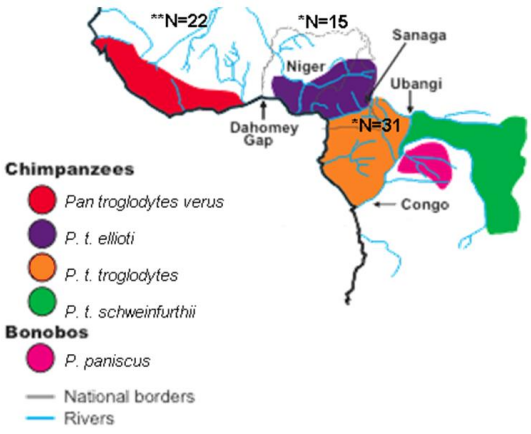
Supplementary Table S5: 2.5 percentile resulting from 100000 coalescent simulations using ms, under three demographic models.

	Constant	**Expansion model**	**Wegmann et al. 2010**
WFDC6(P. t. troglodytes)	-1.13	-1.33	-2.05
WFDC6 (P. troglodytes)	-1.586		
EPPIN(P. t. troglodytes)	-1.00	-1.22	-2.03
EPPIN(P. troglodytes)	-1.589		

Supplementary Table S6: Parameter Estimates and Likelihood Scores under Different Branch models.

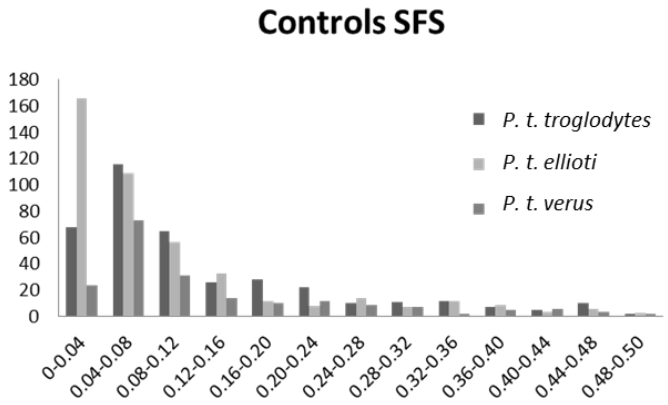
	Parameters for Branches	**Likelihoods**
One ratio	$\omega_{EPPIN\text{-}WFDC6} = 0.4739$	-1388.2014
Two ratios	$\omega_{EPPIN} = 0.4738$	-1387.6820
	$\omega_{WFDC6} = 0.7782$	
Three ratios	$\omega_{EPPIN} = 0.4739$	-1387.5959
	$\omega_{WFDC6others} = 0.7091$	
	$\omega_{WFDC6ancHomoPan} = 1.1656$	
Three ratios	$\omega_{EPPIN} = 0.4738$	-1387.6534
	$\omega_{WFDC6\ others} = 0.8119$	
	$\omega_{WFDC6Pan} = 0.5891$	

Supplementary Figures

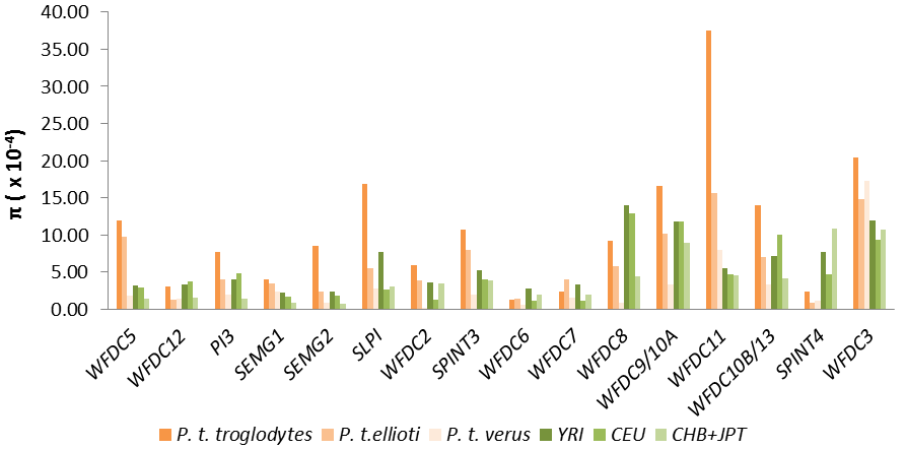


Supplementary Figure S1: Geographic distribution of *P. troglodytes* subspecies in Africa [adapted from Gonder et al. 2011].

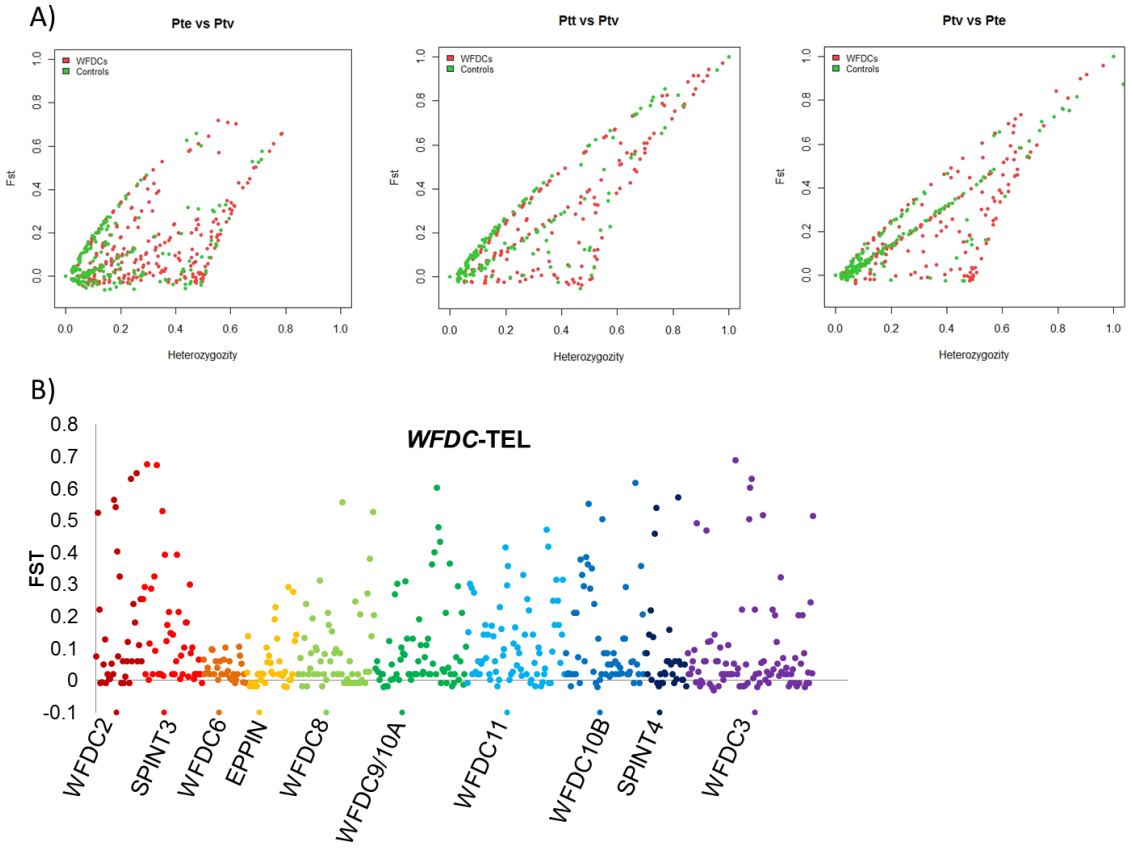
A)



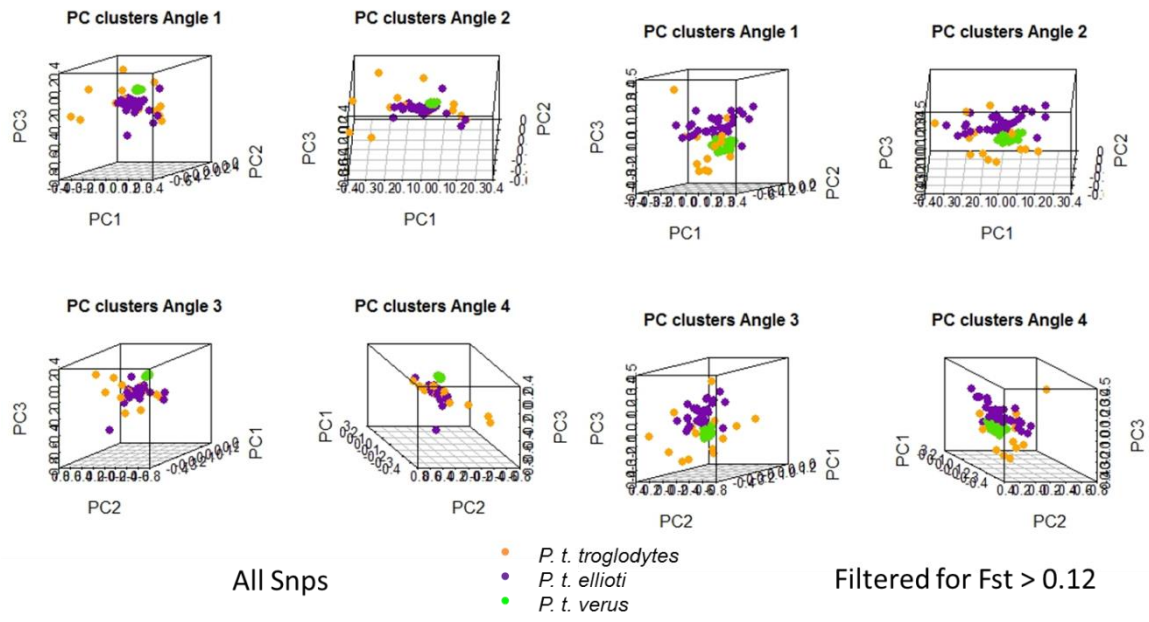
B)



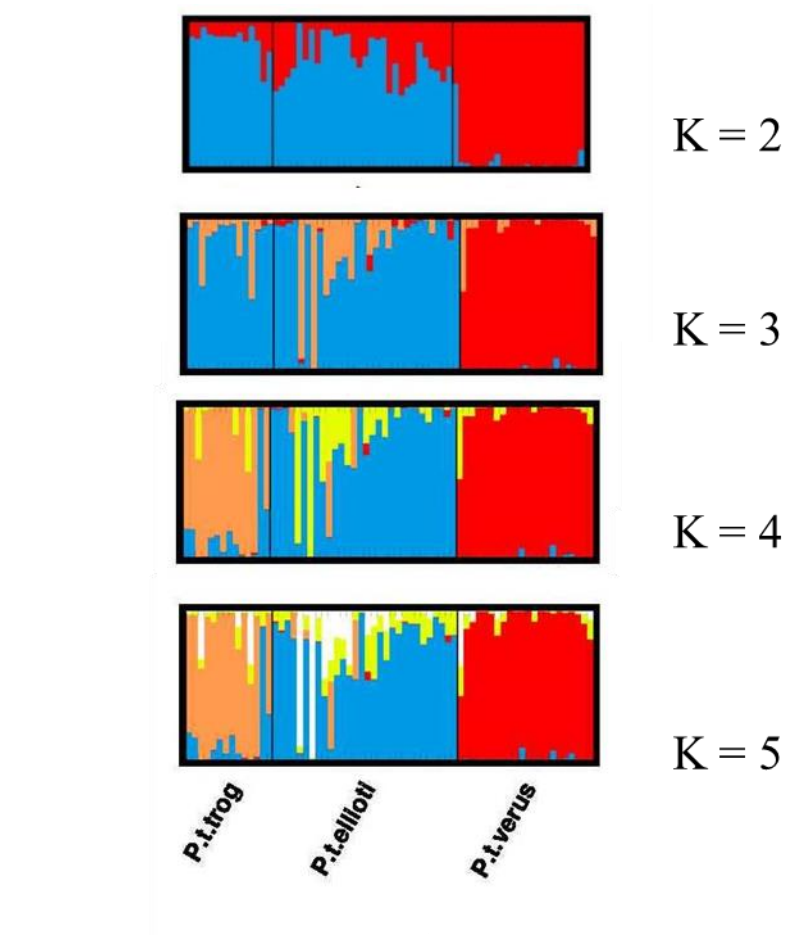
Supplementary Figure S2: A) Folded site frequency spectrum (SFS) in the control regions B) Distribution of nucleotide diversity for all the *WFDC* genes in the sequenced chimpanzee subspecies and the corresponding regions in the 1000 genomes database.

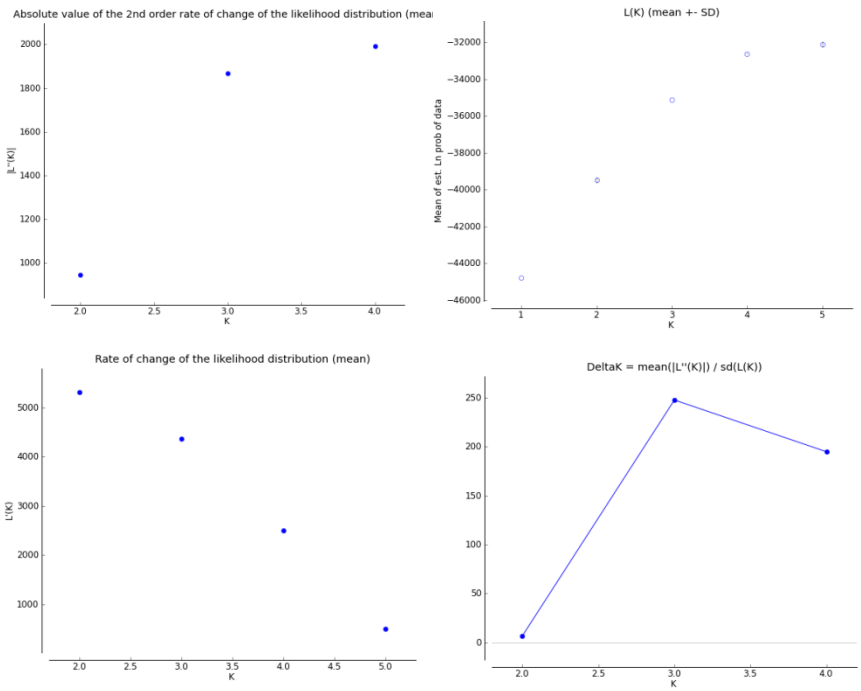


Supplementary Figure S3: A) F_{ST} values in Y axis and heterozygozity in the X axis as calculated in ARLEQUIN. B) F_{ST} values for all the populations, plotted based on genomic position.

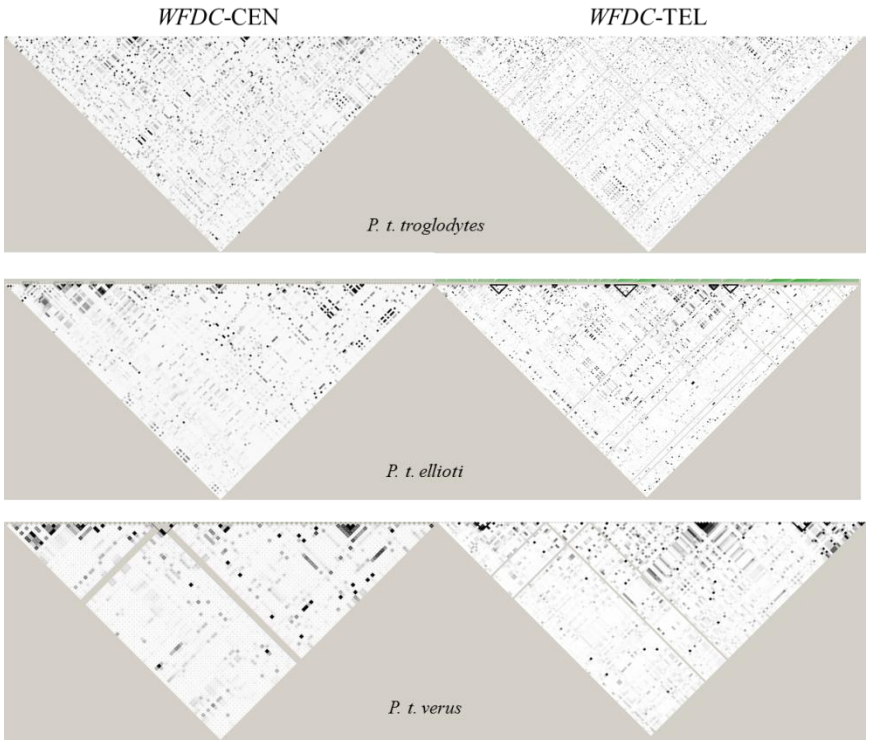


Supplementary Figure S4: Principal component analysis (PCA) plots of markers. Eigenvalues were calculated in eigenstrat and plots were made using R package.

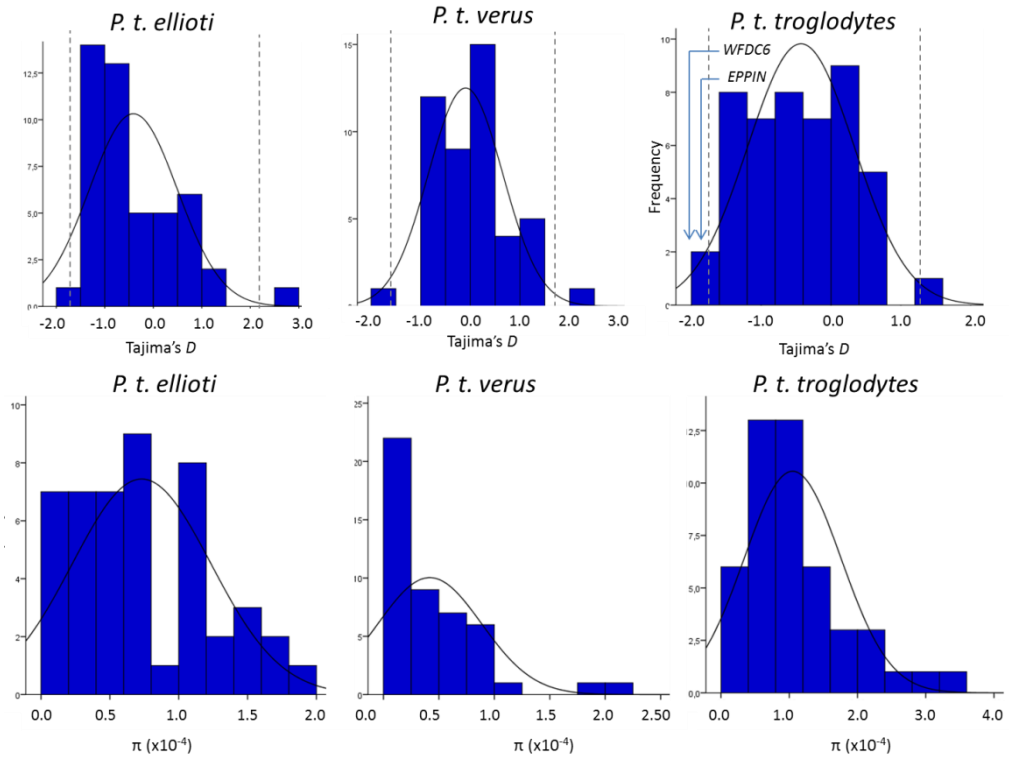




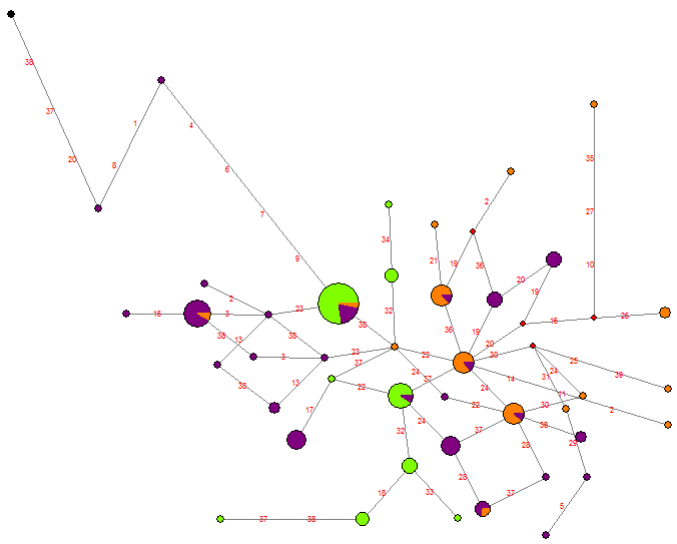
Supplementary Figure S5: A) Bar plot output from STRUCTURE; B) Likelihoods calculated from STRUCTURE outputs.



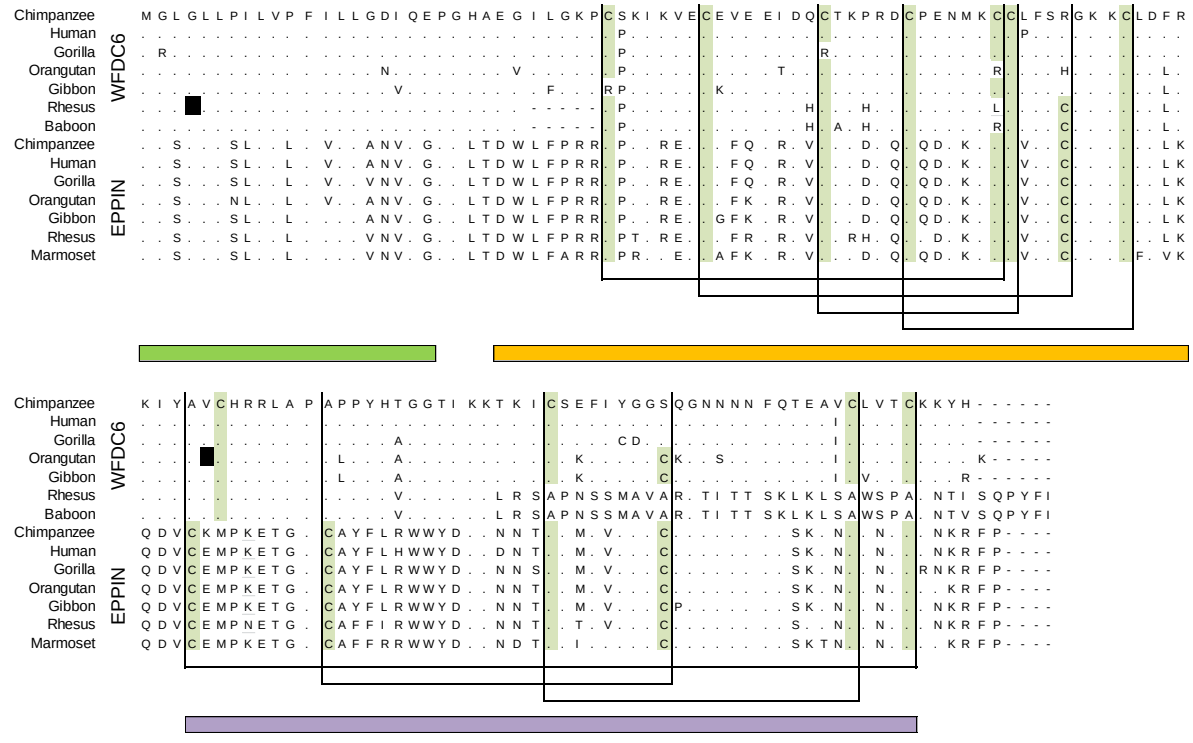
Supplementary Figure S6: LD plots as calculated in Haploview (r^2 statistic).



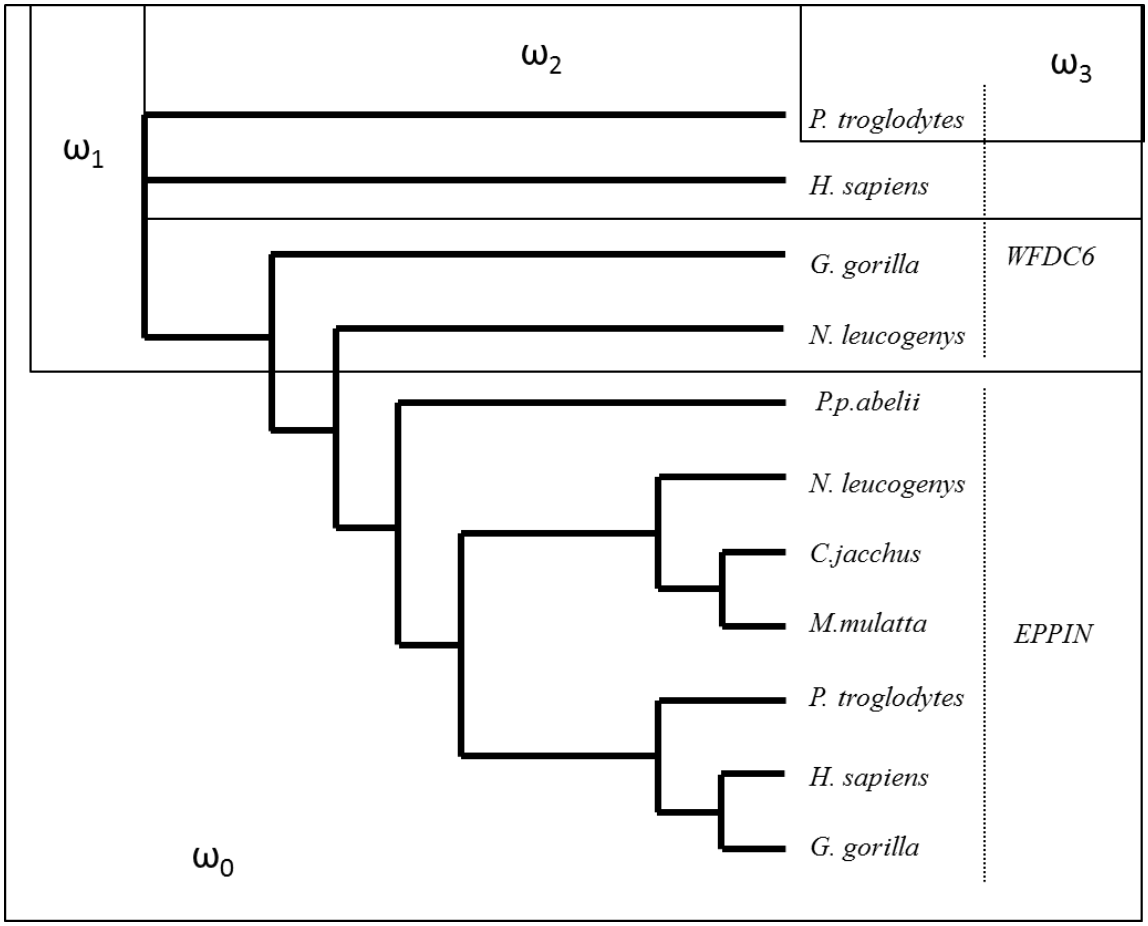
Supplementary Figure S7: Empirical comparisons generated from the 47 control regions. Tajima's D and π were calculated for each regions using SLIDER and plotted using SPSS software. 2.5 and 97.5 percentiles are represented in dashed lines.



Supplementary Figure S8: Inferred *EPPIN*.network haplotypes. Each circle represents a unique haplotype, and its area is proportional to its frequency. Within each circle, *P. t. verus*, *P. t. ellioti*, and *P. t. troglodytes* are labeled in green, purple and orange, respectively. The mutations that differentiate each haplotype are shown along each branch.



Supplementary Figure S9: Amino acid alignment of WFDC6 and EPPIN. Cysteines are marked in light green and disulfide bridges are marked in black lines. Black squares represent Stop Codons.



Supplementary Figure S 10: **Phylogenetic analysis of *WFDC6* and *EPPIN* in primates, showing d_N/d_S ratios. The different branch models are represented as follows: ω_0, one ratio model; ω_1, two-ratio model; ω_2 the three-ratio model 1; and ω_3 the three-ratio model 2.**

Appendix IV

Supplementary Material Article 4

Characterization of the Human *WFDC8: Evolutionary history and* differential allele expression

(in preparation)

Supplementary Tables

Supplementary Table S1: WFDC8 sequence references used for alignment and phylogeny.

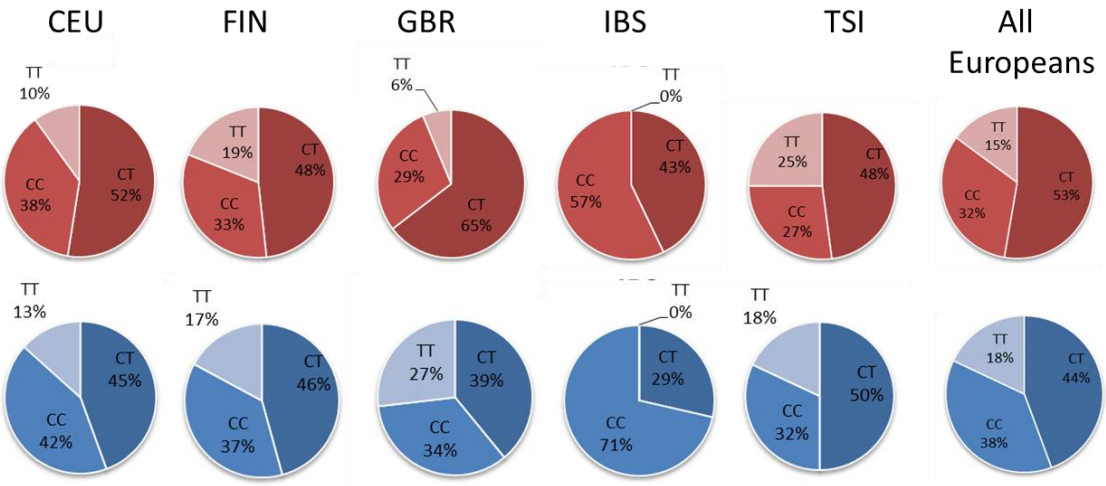
WFDC8	**NCBI reference**
Homo sapiens	NM_130896.2
Pan troglodytes	XM_003316967.1
Pan paniscus	XM_003825972.1
Gorilla gorilla	N/A*
Pongo abelii	XM_002830356.2
Macaca mulatta	XM_001107439.2
Saimiri boliviensis boliviensis	XM_003936446.1
Nomascus leucogenys	XM_003253645.1
Papio anubis	XM_003904689.1
Canis lupus familiaris	XM_534437.2
Mus musculus	NM_001080550.2
Rattus norvegicus	XM_003749629.1

* This sequence was obtained by the BLAT utility at http://genome.ucsc.edu/cgi-bin/hgBlat?command=start

Supplementary Table S2: Samples genotypes.

Sample	DNA	RNA	RNA (post-DNAase I)	Genotype
	(ng/ul)	**(ng/ul)**	**(ng/ul)**	**rs2250860**
14416	78,83	506,37	38.46	G-A
14454	150.6	387,01	33.27	G-A
11667	51,72	454,48	34.85	G-A
876	110,39	573,91	380.81	G-A
7584	15,63	122,07	6.27	AA
2637	51,24	162,35	13.08	AA
17826	88,81	807,11	248.78	G-A
22599	54,67	578,46	4.58	G-A
19638	86,52	613,18	8.49	G-A
11676	33,72	3,04		GG
14873	71,69	497,65	55.17	GG
123805	102.85	48,35	11.55	G-A
22023	76.83	607,20	101.37	G-A

Supplementary Figures



Supplementary Figure S1: Allele frequencies of rs7273669, extracted from the 1000 Genomes Project website.